

EMBODIED, SIMULATION-BASED COGNITION: A HYBRID APPROACH

by

Sean Christopher Kugele

A Dissertation

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

Major: Computer Science

The University of Memphis

May 2023

Copyright© Sean Christopher Kugele

All rights reserved

## **Dedication**

This dissertation is dedicated to the memories of

my father, John Kugele (1944-2007),

who set me on this path,

my research advisor, Stan Franklin (1931-2023),

who showed me how to walk it,

and

my sweet kitten Caramon (2006-2023),

who kept me company every step of the way.

## Acknowledgements

I am forever indebted to Stan Franklin for being my research advisor, career mentor, and friend over many years. Stan stimulated my interest in cognitive science, patiently educated me about the LIDA cognitive architecture, and introduced me to grounded cognition and the work of Lawrence Barsalou—which motivated this dissertation. Our work together will always be one of the high points of my life.

I am also indebted to the members of my committee—Vasile Rus, Andrew Olney, Deepak Venugopal, and Weizi Li. I am especially appreciative of Andrew Olney, for his thought-provoking comments and steadfast dedication to the quality of this dissertation. Furthermore, I would like to give my heartfelt thanks to Weizi Li for agreeing to join my committee in the eleventh hour, when Stan Franklin’s health failed him.

I would like to acknowledge the many contributions of my family and friends. First, I would like to thank the love of my life Tammy Cater for putting up with me and my crazy whims, keeping me well supplied with coffee, and always being emotionally and financially supportive. I would like to thank my mother, Claudia Kugele, for believing in me, unconditionally loving me, and doing everything in her power to make my dreams a reality. I would like to thank my father, John Kugele, for dreaming big, planning well, and setting me on this path many years ago. (I miss you dad.) I would like to thank my “bonus” mother Lynn Kugele for her encouragement, advice, and many calls and text messages reminding me to work on my dissertation. And I would like to thank my good friends Kevin Fitzpatrick and Ann Styrvoky, for keeping me sane through it all with generous gifts and delightful company—

moreover, a very special thank you to Kevin for spending an enormous amount of time reviewing, discussing, and editing this manuscript.

Finally, I would like to thank the past and present members of Cognitive Computing Research Group (CCRG)—in particular, Steve Strain, Pulin Agrawal, Daqi Dong, Javier Snaider, and Ryan McCall—for their many contributions to LIDA and our many enjoyable discussions.

## Abstract

Kugele, Sean Christopher, Ph.D., The University of Memphis, May 2023. Embodied, Simulation-Based Cognition: A Hybrid Approach. Major Professor: Stanley P. Franklin.

Embodied cognition is a paradigm in cognitive science that emphasizes the fundamental role of bodies, environmental interactions, situational contexts, and sensory and motor systems in cognitive processing. Much of the research in embodied cognition has focused on online cognitive activities (such as reactive, overt behaviors) that are directly coupled to environmental stimuli. Such research fails to explain offline cognitive activities (such as planning, deliberation, and mental imagery) that are spatially or temporally decoupled from one's immediate sensory experiences. Embodied, simulation-based theories of cognition attempt to address this shortcoming by proposing a mechanism—modal simulations—by which sensory and motor systems can directly support offline cognitive activities. However, while purely modal, simulation-based theories, such as perceptual symbol systems, have more explanatory power than their purely online counterparts, they also suffer from their own shortcomings. These include largely untenable explanations of abstract concepts and a conspicuous lack of well-specified and well-developed computational models. To address these challenges, I develop a hybrid (modal/amodal) account of embodied, simulation-based cognition based on a neuro-symbolic implementation of the LIDA cognitive architecture. My implementation focuses on (1) grounded, multimodal perception and perceptual learning; (2) action selection and procedural learning; and (3) mental imagery and simulation-based reasoning. This account advances LIDA's conceptual commitments to embodied principles and grounded cognition; contributes to the ongoing scientific discourse on embodied, simulation-based cognition; and advances a

computational framework for realizing simulation-based, autonomous, agential software systems.

## Table of Contents

Chapter	Page
List of Tables .....	xii
List of Figures .....	xiii
1. Introduction.....	1
Perceptual Symbol Systems.....	3
On Abstract Concepts .....	5
On the Need for Computational Accounts of Embodied, Simulation-Based Cognition ....	7
A New Hybrid Account of Embodied, Simulation-Based Cognition.....	8
2. Background.....	12
What is Cognitive Science? .....	13
Mental Representations and Cognitive Processes.....	13
Icons, Indexes, and Symbols.....	16
Symbolic AI and the Symbol Grounding Problem.....	17
Non-Representationalism.....	22
Connectionism .....	26
Hybrid Systems.....	32
Distributional Semantics.....	35
Embodied, Simulation-Based Cognition .....	40
3. A Hybrid Approach to Embodied, Simulation-Based Cognition .....	52



Preliminaries .....	53
Overview .....	59
Mental Representations.....	62
Cognitive Processes .....	71
Amodal Representations for Grounded Cognition .....	75
On Language.....	89
4. Learning Intelligent Decision Agent (LIDA) .....	91
Overview.....	92
The LIDA Cognitive Cycle.....	95
Modes of Action Selection.....	98
The Conscious Learning Hypothesis .....	99
Activation.....	99
Motivations in LIDA.....	100
The Nature of LIDA’s Representations .....	103
The LIDA Conceptual Model and Its Implementations .....	104
5. Grounded Representations, Mental Simulations, and Multimodal Perception.....	106
Background: Variational Autoencoders (VAEs) .....	106
Implementation .....	109
Evaluation .....	129
Discussion.....	136
6. Action-Based Mental Simulation and Motor Cognition.....	139

Background: The Schema Mechanism .....	141
Implementation .....	151
Evaluation .....	173
Discussion .....	184
7. Mental Imagery and Simulation-Based Epistemic Processes .....	186
Mental Imagery .....	187
Kosslyn’s Proto-Model of Visual Perception and Mental Imagery .....	198
Implementation: Mental Imagery in LIDA.....	204
A LIDA-Based Agent .....	228
Related Work .....	241
Discussion .....	250
8. Closing Remarks .....	252
Contributions.....	252
Related Work .....	255
Directions for Future Work.....	261
ES-Hybrid’s Predictions .....	265
9. References.....	267
10. Appendix.....	306
Representational Properties .....	306
Learning Intelligent Decision Agent (LIDA) .....	326
<b>K</b> -Armed Bandit Agent Implementation: Parameters and their Values .....	328

Simulation-Based Agent Implementation:  $\beta$ -VAE Network Architecture ..... 331

## List of Tables

Table 1. Terminological comparison of Drescher's Schema Mechanism and LIDA.....	153
Table 2. Percentage of trails in which agents focused on the best machine (experiment 1) .....	179
Table 3. Comparison between Shanahan's cognitive architecture and LIDA .....	243
Table 4. Comparison between Soar and LIDA.....	249
Table 5. Contributions to the LIDA cognitive architecture. ....	254
Table 6. LIDA's short-term and long-term memory (STM/LTM) modules and codelets. ....	326
Table 7. Implementation parameters and their values ( $k$ -armed bandit agent). ....	328
Table 8. Learned and built-in schemes for $k$ -armed bandit agent ( $k = 8$ ).....	329
Table 9. Learned base-level incentives saliences for $k$ -armed bandit agent ( $k = 8$ ). ....	330

## List of Figures

Figure 1. High-level depiction of Harnad’s hybrid architecture.....	34
Figure 2. Illustration of ES-Hybrid’s major components and their interactions.....	58
Figure 3. ES-Hybrid’s representational system. ....	62
Figure 4. Grounded concept representations as object maps.....	65
Figure 5. Intuitive depiction of direct and indirect grounding.....	70
Figure 6. Depiction of a hypothesized unknown entity. ....	78
Figure 7. Illustration of cognitive indirection. ....	84
Figure 8. LIDA’s cognitive cycle diagram. ....	95
Figure 9. Depiction of a variational autoencoder (VAE).....	107
Figure 10. Sensory Memory and bottom-up perception. ....	110
Figure 11. Grounded representations and conceptual generalization. ....	113
Figure 12. Simulator structure building codelet (SBC). ....	117
Figure 13. Spatial cognitive map. ....	122
Figure 14. A schematic diagram depicting a cognitive “object” map. ....	123
Figure 15. Current activations and resemblance-based best matches.....	132
Figure 16. Stimuli and their mental simulations (Fashion MNIST). ....	136
Figure 17. A synthetic item and its host schema. ....	144
Figure 18. An illustration of a single chain of schemas.....	146
Figure 19. Jeannerod’s covert and overt action stages in LIDA.....	171
Figure 20. State-transition diagram for a modified, k-armed bandit environment. ....	175
Figure 21. Boxplot of experiment 1’s results.....	180

Figure 22. Plots of the first 1500 steps from a single $k$ -armed bandit trial ( $k = 8$ ).....	181
Figure 23. Boxplot of experiment 2's results.....	183
Figure 24. Kosslyn's proto-model of visual perception and mental imagery. ....	199
Figure 25. Visual sensory scene and Perceptual Scene. ....	205
Figure 26. Real and virtual sensory content co-mingle in LIDA's Perceptual Scene. ....	207
Figure 27. Mental imagery and the Perceptual Scene.....	215
Figure 28. Response times in mental rotation across age groups. ....	219
Figure 29. Materials and procedures used in mental rotation experiment. ....	229
Figure 30. Screenshot of the mental rotation experiment's environment. ....	230
Figure 31. Categories of sensory stimuli pairs.....	231
Figure 32. Cognitive cycle diagram for a LIDA-based agent.....	233
Figure 33. Perceptual variance.....	239
Figure 34. Simulation quality and variability. ....	240
Figure 35. High-level comparison of Soar's and LIDA's mental imagery implementations.....	247
Figure 36. Functional mapping of LIDA's module and processes for comparison with Soar....	248
Figure 37. Conceptual depictions of modal and amodal representations. ....	307
Figure 38. Overview of $\beta$ -VAE neural network architecture.....	331
Figure 39. Overview of $\beta$ -VAE encoder. ....	332
Figure 40. Overview of $\beta$ -VAE decoder .....	332
Figure 41. Polyominoes used in mental imagery environment.....	333
Figure 42. All rotations and scales for a single pentomino shape .....	334
Figure 43. A sigmoidal current-activation function.....	335
Figure 44. A heatmap showing current activations. ....	336

Figure 45. Interpolation and extrapolation from latent representations..... 337

## Chapter 1

### Introduction

I date the moment of conception of cognitive science as 11 September, 1956, the second day of a symposium organized by the ‘Special Interest Group in Information Theory’ at [MIT]... The morning began with a paper by Newell and Simon on their ‘logic machine’.  
(Miller, 2003, p. 142)

From its inception, cognitive science<sup>1</sup> and artificial intelligence (AI) have existed in a symbiotic relationship. Practical approaches in AI have inspired new cognitive theories (e.g., the computational theory of mind and connectionism). Similarly, scientific discoveries about the mechanisms of natural minds have inspired new engineering practices (e.g., convolutional neural networks and reinforcement learning). This dissertation exemplifies the mutually beneficial relationship between cognitive science and AI. I design intelligent systems based on cognitive theories, and I attempt to understand the principles and mechanisms governing natural minds by engineering and studying artificial minds.

Cognitive theories are attempts to explicate how minds work. They make claims about the existence and properties of mental representations; the innate or experiential basis of knowledge; the relationship between mind, body, and environment; and the nature of thought and cognitive processes (among other things). This manuscript focuses on a specific class of cognitive theories that I refer to as *embodied, simulation-based cognition* (see Barsalou, 1999,

---

<sup>1</sup> Cognitive science is an inter-disciplinary scientific effort that seeks to understand minds and cognitive processes. Cognitive scientists have traditionally come from psychology, computer science, neuroscience, philosophy, linguistics, and anthropology (see Miller, 2003).



2008, 2010; Bergen, 2012, 2015; Gallese, 2005, 2007; Goldman, 1992, 2012, 2013; Grush, 2004; Jeannerod, 2006; Rizzolatti et al., 1996; Rizzolatti & Sinigaglia, 2016; Shanton & Goldman, 2010; Zwaan, 2004). Embodied, simulation-based theories of cognition fall within the broader paradigm of embodied cognition (EC), which emphasizes an interdependence between minds, bodies, environmental interactions, and situational contexts. A central tenet of EC is that “cognition is for action” (M. Wilson, 2002, p. 632)—the overriding purpose of cognition is the production (i.e., selection and execution) of actions (Franklin, 1995, Chapter 16) and answering the question “What do I do next?”

Research within the EC paradigm has focused on explaining minds using *online* control processes. These processes are said to be directly “coupled” to an environment (e.g., the physical world). Online control processes operate by mapping the current and immediate sensory inputs (environmental stimuli) available to a cognitive system to motor outputs (overt behaviors). These “behavior generating modules” (Brooks, 1990, p. 3) are characterized as being situated<sup>2</sup>, reactive, and requiring little cognitive processing. They are also often described as being non-representational (see Chapter 2). While these perspectives are important, views of cognition that only include online control processes fail to explain cognitive activities that seem to be “detached” from immediate sensory inputs and motor activity. These activities include (but are not limited to) planning, reasoning, introspection, problem solving, and mental imagery, as well

---

<sup>2</sup> *Situated cognition* is a school of thought associated with the embodied cognition paradigm. Roth and Jornet (2013) stated that the central hypothesis of situated cognition is that intelligent behavior arises from a dynamic coupling between individuals and their environments rather than from their minds (e.g., brains) only. According to this view, information does not exist in minds prior to environment interactions, but, instead, emerges from those interactions.

as more pedestrian ones like the recall of long-term memories and daydreaming. In these cases, *offline* control processes appear to be needed.

Embodied, simulation-based theories of cognition attempt to address this shortcoming by proposing a mechanism by which offline control processes can take on an embodied and “situated” character. That mechanism is *sensorimotor simulations*: mental simulations<sup>3</sup> based on the (re-)activation of sensorimotor systems and their *grounded* (see Chapters 2 and 8) representational derivatives. In other words, the cognitive subsystems most directly responsible for interfacing with the world are recruited when reasoning about the world.

### **Perceptual Symbol Systems**

Barsalou’s (1999, 2008, 2010) theory of grounded cognition—known as Perceptual Symbol Systems (PSS)—is arguably the most influential and well-developed theory of embodied, simulation-based cognition. According to this theory, the patterns of (neural) activation that occur in sensorimotor systems during perception and action can be learned into long-term memory (albeit in a partial and attenuated form). If later recalled from long-term memory, the associated (re-)activations of sensorimotor systems can function as *perceptual symbols*: grounded mental proxies for entities, objects, and events in the world.

Barsalou (1999) characterized perceptual symbols as being *analogical* and *modal*.<sup>4</sup> They are analogical because they share properties with, or in some way resemble, their originating

---

<sup>3</sup> Consciously accessed mental simulations are referred to as “mental images,” and the cognitive processes associated with them are collectively referred to as “mental imagery” (see Kosslyn, 1994; Kosslyn et al., 2006). Mental imagery and imagistic processes are the topics of Chapter 7.

<sup>4</sup> These representational properties are discussed in detail in the Appendix.

sensory and perceptual (mental) states. They are modal because “they are represented in the same systems as the perceptual states that produced them” (Barsalou, 1999, p. 578). Moreover, they are often multimodal, accounting for content originating in multiple sensory and motor modalities.

Perceptual symbols that correspond to instances of the same concepts (e.g., categories of things) can eventually become integrated into *simulators*. Simulators are described by Barsalou as combinations of generative processes, empirical knowledge, and genetic predispositions that allow an individual to adequately represent concepts during offline cognition (Barsalou, 1999, sec. 2.4.3). Specifically, simulators generate (modal) mental simulations of the concepts they represent. These mental simulations can then be integrated with other mental simulations to form rich, contextualized, virtual scenes (i.e., situated conceptualizations; Barsalou, 2016b). Critically, the same perceptual processes that operate on external environmental stimuli must be able to operate on these internal virtual scenes.

According to Barsalou’s (1999) theory, grounded representations and mental simulations are sufficient to implement a “fully functional conceptual system” without the need for *amodal* (ungrounded) symbols, such as those used in classical symbolic AI and related cognitive theories (e.g., the computational theory of mind [CTM] and language of thought hypothesis [LOTH]; see Chapter 2). Specifically, Barsalou (1999) claimed that a perceptual symbol system can, in theory, support the following cognitive functions: the distinction between *types* (i.e., categories) and *tokens* (i.e., instances of categories); *categorical inference* (i.e., assigning categorical membership to instances); *combinatorial and recursive productivity* (i.e., binding and nesting concepts within concepts to generate new conceptual structures); *propositions*; and *abstract*

*concepts*. While Barsalou (1999) provided an intriguing, albeit extremely high-level argument to support each of these, his account of abstract concepts is arguably the weakest aspect of his theory. Consequently, it has been the subject of a great deal of criticism and controversy.

### **On Abstract Concepts**

Abstract concepts are notoriously difficult to explain in terms of embodied principles. Concrete concepts (such as chairs, cats, and bicycles) have “bounded, identifiable referents that can be perceived with our senses” (Borghetti et al., 2017, p. 1). In contrast, abstract concepts (such as truth, time, and transfinite numbers) lack bounded, perceivable referents. Given these definitions, Barsalou’s (1999) purely perceptual and empirical account of abstract concepts meets immediate resistance.

Briefly, Barsalou contended that abstract concepts are always grounded (see Barsalou, 1999, p. 577) and can be learned via three cognitive mechanisms—framing, selectivity, and the use of introspective perceptual symbols (see Barsalou, 1999, p. 600). He described this empirical learning process as follows: First, abstract concepts are framed against a simulated, background event sequence. Second, selective attention identifies the content corresponding to those abstract concepts against their simulated backgrounds. Finally, introspective states and proprioceptive events are incorporated with the perceptual symbols for those abstract concepts, eventually culminating in simulators for those concepts.

Unfortunately, this idea seems flawed in principle, regardless of the specific perceptual mechanisms involved. As one peer commentator observed, “abstract concepts are not typically associated with any particular event sequences or introspections” (Ohlsson, 1999, p. 631). In another peer commentary, Toomela stated,

[Barsalou's] perceptual theory of knowledge is not very convincing regarding abstract concepts... [because] there exists a kind of knowledge that must be amodal in essence: the knowledge about a world which is qualitatively out of reach of our senses. Humans do not possess perceptual mechanisms for perceiving electromagnetic fields, [for example]... How such knowledge is constructed is not explained in Barsalou's theory. (Toomela, 1999, p. 633)

Other peer commentaries, such as one by Adams and Campbell, also concluded that Barsalou's account of abstract concepts fails because—by definition—there are no perceivable exemplars for abstract concepts. For example, how does one simulate and perceive the differences between “chiliagons” (1000-sided polygons) and “myriagons” (10,000-sided polygons)? How does one simulate or perceive “the infinity of parallel lines in a Lobachevskian space” (Adams & Campbell, 1999, p. 610)? These concepts appear to depend on amodal definitions. Similar arguments have led others to contend that abstract and concrete concepts are *different in kind*, requiring different representational frameworks (Crutch & Warrington, 2005; Dove, 2009), and there is some neuroscientific evidence that has been interpreted in support of this claim (e.g., see Binder et al., 2005; Desai et al., 2018; Wang et al., 2010).

Barsalou's more recent work has attempted to shift the focus away from the concrete versus abstract conceptual dichotomy. For example, Barsalou et al. (2018) stated, “we increasingly doubt whether terms like ‘concrete’ and ‘abstract’ are ultimately useful and informative in describing concepts” (Barsalou et al., 2018, p. 5), and they suggest replacing this distinction with other conceptual dichotomies (e.g., “external” versus “internal” situational elements, and situational “elements” versus situational “integrations.”) While these ideas may

prove to be useful, they do not obviate the need to explain the nature of concrete and abstract concepts, or the need to understand their acquisition and processing within cognitive systems.

Arguably, Lakoff and Johnson (1999, 1980/2008) provided the most compelling embodied account of abstract concepts. They suggested that we think about abstract concepts through metaphors and analogies with concrete concepts. Thinking of time in terms of motion (“time *flies* like an arrow”), or similarity in terms of physical proximity (“orange is *closer* in color to red than blue”), or categories in terms of physical containers (“humans are primates *in* the genus Homo”) are examples of this metaphorical mode of thought.

While Lakoff and Johnson are undoubtedly correct—humans often use analogical thinking when grappling with abstraction—this is still not the whole story. Humans also have a capacity to learn, use, and communicate abstract concepts in the absence of relevant concrete metaphors. This frequently happens in abstract domains such as mathematics, where many concepts can only be adequately expressed by their relationship with other abstract concepts. If analogies between abstract and concrete concepts do emerge, they often reflect deep insights that only come from rare genius or following extensive experience (such as those of a professional mathematician). A more expansive explanation of abstract concepts is needed.

### **On the Need for Computational Accounts of Embodied, Simulation-Based Cognition**

Abstract concepts aside, Barsalou’s theory of perceptual symbol systems is arguably the most comprehensive and compelling simulation-based cognitive theory to date. And yet, despite its conceptual appeal and the growing neuroscientific and experimental support for many of its assertions (see Barsalou, 2008), the computational mechanisms for implementing Barsalou’s vision have remained elusive. Barsalou (2009) stated,

Perhaps the most pressing issue surrounding this area of work is the lack of well-specified computational accounts. Our understanding of simulators, simulations, situated conceptualizations and pattern completion inference would be much deeper if computational accounts specified the underlying mechanisms. (Barsalou, 2009, p. 1287)

Pezzulo et al. (2013) reiterated this urgent need for computational accounts when they stated,

Despite its growing popularity, the full potential of [modal, grounded cognition] has not yet been demonstrated; and this is not only a matter of obtaining new empirical demonstrations of the importance of grounding for cognition. The framework is empirically well-established, but the theories are relatively underspecified. *A real breakthrough might result from the realization of explicit computational models that implement grounding in sensory, motor and affective processes as intrinsic to cognition* [emphasis added]. (Pezzulo et al., 2013, p. 2)

Unlike the computational theory of mind, which has well-established mechanisms in classical symbolic AI and connectionist systems (see Chapter 2), theories of embodied, simulation-based cognition are still searching for their computational footing.

### **A New Hybrid Account of Embodied, Simulation-Based Cognition**

To address these challenges, I develop a new *hybrid* (modal/amodal) account of embodied, simulation-based cognition based on a *neuro-symbolic* implementation of the LIDA cognitive architecture (Franklin et al., 2016). Neuro-symbolic systems combine neural networks and symbolic AI. A sentiment shared by many researchers (including myself) is that these systems

have the potential to be more robust, transparent, interpretable, and capable than techniques based solely on connectionist or symbolic AI alone (e.g., see Garcez & Lamb, 2020; Mao et al., 2019; Marcus, 2020; Sarker et al., 2021).

One common neural-symbolic design pattern uses neural networks to transform non-symbolic inputs (e.g., images) into symbolic representations (e.g., words) which are then manipulated by a symbolic reasoning system. Kautz (2022) referred to this as a “**Neuro | Symbolic** system”<sup>5</sup>. The neuro-symbolic implementation presented here could be seen as a variant of this basic design; however, it also deviates from that design by employing a *non-symbolic* (imagistic) reasoning system (see Chapter 7), and a combined symbolic and non-symbolic associative memory that is grounded in a multi-dimensional latent vector space (see Chapter 5).

Throughout the remainder of this manuscript, I will refer to the cognitive theory developed here as ES-Hybrid (Embodied Simulation-Hybrid). ES-Hybrid draws a great deal of inspiration from Barsalou’s theory of Perceptual Symbol Systems (PSS). Indeed, it developed out of an initial effort to implement a perceptual symbol system within the LIDA (Learning Intelligent Decision Agent; Franklin et al., 2016) cognitive architecture; however, in doing so, it became clear that Barsalou’s theory could only serve as a partial account of cognition. In particular, PSS’s inability to provide an adequate account for abstract concepts is indicative of a deeper theoretical limitation. Unfortunately, this issue cannot be addressed within the confines of

---

<sup>5</sup> Kautz’s characterization of **Neuro | Symbolic** systems is consistent with the hybrid (symbolic/connectionist) architecture suggested by Harnad (1990) as a solution to the symbol grounding problem. Both this architecture and the symbol grounding problem are discussed in Chapter 2.



that theory because Barsalou expressly eliminates that which is needed: amodal representations. Addressing this issue requires a hybrid (modal/amodal) account of cognition. Furthermore, many researchers (other than Barsalou) have made important contributions to embodied, simulation-based cognition (both experimentally and theoretically). For example, Jeannerod's "theory of motor cognition" (see Jeannerod, 1995, 2001, 2006) was particularly influential in the implementations of action selection and action-based mental imagery developed here (see Chapters 6 and 7).

While there are many aspects of ES-Hybrid that need to be addressed, the work here focuses on several foundational conceptual and computational issues:

- (1) grounded, multimodal representations and bottom-up perception,
- (2) an implementation of action selection and procedural memory compatible with simulation-based theories of cognition,
- (3) the fundamental operations of mental imagery and imagistic cognitive processes, and
- (4) a more complete account of abstract concepts.

Other contributions include advancing LIDA's implementations of procedural learning, motivational learning, action selection, grounded and ungrounded representations, (preconscious and never conscious) mental simulations, and (conscious) mental imagery. Many of these contributions are currently limited to high- or mid-level designs; however, several low-level designs and software implementations are also developed. For example, software implementations that utilize  $\beta$ -variational autoencoders (Higgins et al., 2017) and a modified version of Drescher's (1991) Schema Mechanism.

The remainder of this manuscript has the following structure: Chapter 2 surveys the cognitive theories and the “symbol grounding problem” (Harnad, 1990) that motivated the development of embodied, simulation-based cognition. Chapter 3 provides an overview of ES-Hybrid, including its mental representations and cognitive processes. Chapter 4 provides background on the LIDA cognitive architecture (Franklin et al., 2016), including its cognitive cycle, modules and processes, modes of action selection, and the conscious learning hypothesis. The next three chapters develop specific ES-Hybrid functionality within a neuro-symbolic implementation of LIDA: Multimodal perception, mental simulation, and grounded, modal representations are developed in Chapter 5; action-based mental simulation and motor cognition is developed in Chapter 6; and mental imagery and imagistic processes are the topic of Chapter 7. Chapter 8 concludes the manuscript with closing remarks and directions for future work.

## Chapter 2

### Background

if the meanings of symbols in a symbol system are extrinsic, rather than intrinsic like the meanings in our heads, then they are not a viable model for the meanings in our heads:

Cognition cannot be just symbol manipulation. (Harnad, 1990, p. 339)

While this manuscript focuses on a class of cognitive theories that I refer to as *embodied, simulation-based cognition*, these theories did not develop in a vacuum. To fully appreciate their significance, they need to be examined with respect to the historical contexts that shaped and motivated their development. Moreover, since cognitive theories often develop in response to the perceived failings of their predecessors, they are often best understood *in contrast to* their predecessors.

With this in mind, I devote this chapter to outlining the cognitive theories and challenges (e.g., symbol grounding) that gave rise to embodied, simulation-based theories of cognition. In so doing, I introduce terminology and concepts that will be used throughout this manuscript. Moreover, many of the ideas presented in this chapter have been incorporated into ES-Hybrid—the hybrid account of embodied, simulation-based cognition developed throughout this text. Therefore, a basic familiarity with these ideas will be beneficial for understanding the chapters that follow. Whenever possible, cognitive theories are presented alongside pertinent engineering approaches that illustrate aspects of those theories in practice.

## **What is Cognitive Science?**

Cognitive science emerged in the 1950s in response to the perceived failings of behaviorism. Behaviorism attempted to establish psychology as an empirical, natural science (akin to physics and chemistry) by focusing exclusively on the relationship between observable behaviors and their antecedent environment stimuli. Minds and mental phenomena, such as consciousness and mental states, were largely treated as pseudo-scientific concepts. John Watson, one of the chief architects of psychological behaviorism, envisioned experimental psychology as a science that shunned terms such as consciousness, mental states, mind, mental content, and mental imagery (Watson, 1913, p. 166), and was instead focused on explicating animal behaviors in terms such as stimulus and response, habit formation, and habit integrations (Watson, 1913, p. 167).

Behaviorism dominated psychological thought for the first half of the twentieth century, but this approach proved to be too restrictive and eventually gave way to the “cognitive revolution.” Miller (2003) described the cognitive revolution as a “counter-revolution” that “brought the mind back into experimental psychology” (Miller, 2003, p. 142). This interdisciplinary movement eventually gave rise to the field of study known as *cognitive science*—an approach to understanding minds and behaviors in terms that include mental phenomena in addition to environmental stimuli and situational (e.g., physical, social, cultural) contexts.

## **Mental Representations and Cognitive Processes**

Thagard (2020) stated that “the central hypothesis of cognitive science is that thinking can best be understood in terms of representational structures in the mind and computational procedures that operate on those structures.” In my opinion, this characterization is only partially accurate. While these concepts are fundamental to many cognitive theories, some theories are *non-*

*representational* (see Braitenberg, 1986; Brooks, 1991b; Dreyfus, 2002; Gallagher, 2008, 2017) and others intentionally avoid computational analogies (Port & Van Gelder, 1995; see Thelen & Smith, 1994). That said, a cognitive theory’s account of mental representations and cognitive processes—even if that account ultimately denies their existence—is a useful dimension for understanding and categorizing most, if not all, cognitive theories. Therefore, understanding what is meant by “mental representations” and “cognitive processes” is requisite knowledge for understanding cognitive science and its theories.

According to many cognitive theories (including the computational theory of mind and connectionism, which are described later in this chapter), *mental representations* are the means by which information and knowledge are represented within a cognitive system (e.g., a nervous system). Mental representations are often said to be internal (mental) states that are “about”<sup>1</sup> the objects, situations, and events (etc.) that one’s thoughts are directed towards. In other words, mental representations stand-in for the things they refer (i.e., their referents), and they can serve as proxies for those referents when thinking about them in their absence (e.g., during offline cognition). Mental representations are commonly compared to the data structures (graphs, trees, linked lists, arrays, pixel maps, etc.) that are used to arrange and depict data within a computer.

*Cognitive processes*, on the other hand, are “control structures” (Newell, 1973) that implement and/or support the mental capabilities hypothesized to exist in natural minds; these might include perception, attention, action selection, motor control, and mental simulation. Collectively, these control processes govern how minds operate within a given situational

---

<sup>1</sup> In the philosophical literature, this capacity for a mental state to be about, or stand in for, something is referred to as its “intentionality” (see Brentano, 1874/2012; Jacob, 2020; Searle, 1980).

context. Newell (1973) explained the idea of a control structure through a computer programming analogy. Specifically, he described them as input-dependent sequences of computational instructions (such as function calls and input/output operations) that are used to implement mental competences. More generally, control structures can be defined as those mechanisms that enable *autonomous agents* (see Franklin & Graesser, 1997) to answer the question, “What do I do next?” Given this broader definition, *minds* can be conceptualized as the “control structures of autonomous agents” (Franklin, 1995, p. 412).

Mental representations cannot be adequately understood in isolation of the cognitive processes that operate on them<sup>2</sup>. By analogy: records require record players, CDs require CD players, and MP3s require media players (with an MP3 codec). While records, CDs, and MP3s are capable of representing the same *contents*—for example, the same song—their representational *formats* are largely different and incompatible with other devices. In other words, it is only the right combination of representational medium (e.g., a record) and consuming device (e.g., a record player) that will produce the intended result: in this case, music. Similarly, in the absence of an appropriate cognitive process to operate on them, mental representations are inert and incomprehensible within a cognitive system. Therefore, cognitive theories that posit the existence of mental representations are obligated to detail the cognitive processes that operate on them.

---

<sup>2</sup> See Anderson (1978) for a more detailed discussion about the interdependence between representations and processes, and its significance in evaluating cognitive theories.

## Icons, Indexes, and Symbols

Semiotics is the study of signs, sign processes, and the production of meaning. Loosely speaking, a *sign* is anything (e.g., words, events, visual depictions, physical objects) that conveys the “meaning” of something (other than itself) to someone (an interpreter of that sign). Therefore, semiotics broadly encompasses the study of anything that intentionally or unintentionally “stands for” something else in the mind of an interpreter (see Chandler, 2022). Consequently, semiotics can be used to characterize mental representations and cognitive processes: mental representations are (internal) signs that signify their (external and internal) referents for the benefit of an interpreting cognitive process. Indeed, Peirce—a founder of the field of semiotics—believed that all thoughts could be understood in terms of mental signs (see Peirce, 1893–1913/1998, p. 10).

Peirce’s theory of semiotics includes three types of signs—icons, indexes, and symbols (Peirce, 1867–1893/1992, pp. 225–228). *Icons* signify their referents by resemblance. That is, they are “likenesses” of their referents, sharing one or more distinctive qualities with them. For example, a drawing of a chair can be used as an icon for the concept of “chairs” (or some particular chair). Geometric diagrams and topographical maps are other common iconic signs. Crucially, an icon’s form must convey enough informational content to an observer for them to recognize the thing being signified.

*Indexes* signify their referents through observable or inferable connections with their referents; in other words, indexes imply the existence of their referents by their very presence. Smoke can be an index of fire, yelling can be an index of anger, and pain can be an index of an

injury. Indexes “point to” their referents’ existence, drawing attention to them like fingers pointing to objects in a scene.

Finally, *symbols* signify their referents through social or cultural conventions. For example, mathematicians in the eighteenth century established the convention of using the symbol  $\pi$  to refer to the ratio of a circle’s circumference to its diameter. There is nothing inherent in  $\pi$ ’s form to indicate its relationship to that concept, and any other symbol could have served this purpose. As another example, consider the English word “chair.” The word CHAIR is a symbol that could refer to “an object designed to be sat on” or, alternately, “a person with a particular position of authority or prestige in an organization.” In this case, the sign’s meaning is ambiguous in the absence of an appropriate linguistic or situational context. Furthermore, SILLA (Spanish), STOL (Danish), and 椅子 (Chinese) are linguistic symbols that are also used to refer to chair-like objects, but they are based on different (linguistic) conventions.

### **Symbolic AI and the Symbol Grounding Problem**

Symbolic AI refers to any approach to engineering intelligent systems that is primarily based on the explicit, rule-based manipulation of *symbolic representations*. In their most basic form, symbolic representations (such as tokens and variables) are atomic (i.e., irreducible) representations that can be used to signify any imaginable concept, object, entity, or situation. However, symbolic representations can also be organized into composite (or structured) representations that contain two or more distinct symbols. Simple examples include mathematical relations ( $x > y$ ), logical statements ( $P \rightarrow Q$ ), and predicates (IS\_SISTER(ALICE, BOB)). More complex examples include semantic networks (see Sowa, 1991/2014), frames (see Minsky, 1975), and ontologies (see Gruber, 1995).



By applying various rule-based computations to manipulate these symbolic representations (such as, mathematical or logical operations), symbolic AI systems are able to make inferences, answer queries, and solve a variety of problems. Built-in symbolic knowledge<sup>3</sup> can be represented using propositions or statements in first-order logic. And symbolic reasoning can be implemented using search algorithms and clever heuristics (e.g., see Coulom, 2006 [Monte Carlo Tree Search]; Hart et al., 1968 [A\*]; Newell & Simon, 1961 [means-ends analysis]).

Based on the initial successes<sup>4</sup> of these approaches (particularly in abstract problem domains such as strategy games, logic, and theorem proving), Newell and Simon (1976) conjectured that *any* system exhibiting “general intelligence”<sup>5</sup> will *necessarily* be based on symbol manipulation. This hypothesis is known as the “physical symbol system hypothesis” (PSSH). Note that the PSSH is not merely a statement about engineering intelligent software. It is a claim that artificial *and* natural systems exhibiting general intelligence *must* be based on symbolic manipulation. The PSSH, along with the language of thought hypothesis (LOTH; Fodor, 1975, 2008), helped establish the computational theory of mind (CTM; see Rescorla,

---

<sup>3</sup> The Cyc project, started in 1984, is a long-running attempt at hand-engineering “common sense” in software to facilitate the construction of symbolic *expert systems*. As of this writing, Cyc’s knowledge base is said to contain “10,000 predicates, millions of collections and concepts, and more than 25 million assertions” (*Cyc’s Knowledge Base – Cycorp Inc.*, n.d.). According to Cycorp, it has taken over *4 million hours* to develop this knowledge store and its associated inference engine.

<sup>4</sup> These successes included the “Logic Theorist” (Newell et al., 1957) and the “General Problem Solver” (Ernst & Newell, 1969), which were the basis for the Soar cognitive architecture (Laird, 2012).

<sup>5</sup> Newell and Simon (1976) characterized “general intelligence” as the ability to perform actions that show the same “scope of intelligence” as human actions, are “appropriate to the ends of the system,” and are “adaptive to the demands of the environment... within some limits of speed and complexity” (p. 116). Other definitions of general intelligence include one from Goertzel and Pennachin (2007) who stated that “[a] general intelligence must be able to carry out a variety of different tasks in a variety of different contexts, generalizing knowledge from one context to another, and building up a context and task independent pragmatic understanding of itself and the world” (p. 74).

2020) as a dominant cognitive theory. This theory holds that minds reason and effect the execution of intelligent behaviors through the rule-based manipulation of symbolic mental representations.

### ***The Symbol Grounding Problem***

An issue that can occur with purely symbolic approaches to AI and related theories of mind is that it is not obvious how (internal) symbols can be meaningfully connected to the objects, entities, and events in a non-symbolic (external) environment based solely on their relationship with other (internal) symbols. Recall that the connection between symbols and their referents are based on pre-assigned conventions and their associated, bounding contexts. Unless those conventions and contexts are built into a system a priori, how can they be established?

This issue is often glossed over in practice because humans are generally “in the loop” to interpret the symbolic results of a machine’s computational efforts, effectively connecting symbols with their meanings exogenously and after the fact. Harnad (1990) illustrates this *symbol grounding problem* by offering, as an example, the formidable task of trying to learn Chinese as a second language (or more appropriately, a first language) when the only information at your disposal is a Chinese-to-Chinese dictionary: “[using] the dictionary would amount to a merry-go-round, passing endlessly from one meaningless symbol... to another... never coming to a halt on what anything meant” (Harnad, 1990, p. 339).

Harnad’s argument is reminiscent of Searle’s (1980) Chinese room thought-experiment, which he used to argue against the possibility of Strong AI (artificial general intelligence) and machine understanding based on symbolic manipulation. Searle’s thought experiment asks us to imagine Searle locked in a little room with several stacks of Chinese documents and a book

written in English. Searle is unable to read Chinese, so the Chinese documents are completely incomprehensible to him. However, he does know English, so he can read and comprehend the contents of the book. The book (which corresponds to the room's current "program") contains a set of instructions that describe how to map the characters in the Chinese documents (i.e., the room's "inputs") to corresponding responses. While Searle does not understand Chinese, he can recognize the forms of the Chinese characters in the documents well enough to find the appropriate entry in the English rule book. Therefore, using this rule book, Searle can produce the correct "output" symbols (also in Chinese) based on whatever the rule book was intended to do by its author (i.e., a programmer).

From the perspective of an outside observer, it may appear like the room understands Chinese. However, despite this outward appearance of understanding, Searle argued that it is an illusion since the processes responsible for these symbolic manipulations lack *intentionality*. That is, with respect to the "mind" of the room (i.e., Searle and his instruction book), the Chinese symbols are not directed towards (i.e., about) the objects and states of the world that they represent (Searle, 1980, p. 424). Instead, they are only directed internally to more symbols, and this disconnect cannot be resolved intrinsically by the system.

According to Searle, the Chinese Room described above operates on the same principles as computers running symbolic AI programs. Therefore, Searle concluded that understanding was *extrinsic* to purely symbolic systems. The meaning (intentionality) of their symbols (inputs and outputs) and the purposes behind their symbolic manipulations exist entirely outside of those systems. In particular, their meaning depends on the interpretations provided by the users of those systems and the intentions of those that programmed them. In Harnad's words, meaning is

“parasitic on the fact that the symbols have meaning for *us*” (Harnad, 1990, p. 339). Solving this disconnect between symbols and their referents *intrinsically* rather than extrinsically is the essence of the symbol grounding problem.

### ***Nativism and The Language of Thought Hypothesis***

The language of thought hypothesis (LOTH; Fodor, 1975, 2008) is a cognitive theory that falls within the computational theory of mind and, by extension, symbolic AI. The LOTH argues that thoughts are represented using an *innate*, private mental language often referred to as Mentalese. Fodor (1975, 2008) compared Mentalese to a built-in “machine language” that contains the symbolic primitives necessary for thought. Like acquired, natural (i.e., spoken and written) languages, Mentalese is assumed to have syntax and compositional semantics. Mentalese “words” correspond to worldly concepts (such as, cats, zebras, umbrellas, and liberty) that can be structured into complex expressions that depend on the semantic properties of their parts. Fodor (1975, 2008) further suggested that these *unlearned* conceptual symbols could be connected to worldly experiences using a compilation/decompilation-like translation process. The purpose of this process is to map to and from Mentalese and the concepts, objects, entities, situations, and events that Mentalese symbols and expressions refer in the world<sup>6</sup>. The LOTH advocates for the view that it is possible, in theory and in practice, to create a formal symbol system with an intrinsic understanding of worldly concepts by virtue of a *fixed, built-in set of symbols* and a *translation process*. And it is the role of the translation process to connect internal Mentalese symbols to their corresponding referents in the external world. Unfortunately, implementing such

---

<sup>6</sup> Proponents of the LOTH are careful to note that Mentalese does not include a 1-to-1 mapping between the apparently unlimited variety of possible worldly concepts and Mentalese symbols. Instead, Mentalese consists of a static, finite set of internal symbols that can be *productively combined* to represent all worldly phenomena.

a built-in translation process for a complex or realistic environment seems all but impossible, since this process must be able to express the worldly “essence” of every conceivable concept in Mentalese, and simultaneously recognize their sensory signatures in the world. All of this *without the benefit of worldly experiences*. As such, its existence in natural systems seems both implausible and unverifiable. As a result, the LOTH has been widely criticized (Rescorla, 2019), and by Fodor’s own admission, his account of concept acquisition is underdeveloped (see Fodor, 2008, sec. 5.4).

In general, purely symbolic approaches (like LOTH) tend to be excellent at implementing *offline* cognitive processes such as problem solving, planning, and deliberation, but perform poorly on complex perceptual tasks (e.g., object recognition) and tasks requiring fine motor control. Symbolic AI requires a separate translation (transduction) process to map non-symbolic inputs (environmental stimuli) to internal symbols, and internal symbols to non-symbolic outputs (motor commands). Explicating the cognitive processes that exist at the interface between symbolic reasoning systems and their environments is arguably symbolic AI’s greatest challenge.

### **Non-Representationalism**

One strategy for solving the symbol grounding problem is to do away with representations entirely. In such systems, the symbol grounding problem is irrelevant since there are no symbols. However, the more general issues of explicating intrinsic meaning still apply. This *non-representational* approach was championed (from a computational perspective) by Rodney Brooks in the 1980s and 1990s.

Brooks (1990) argued that the “physical symbol system hypothesis” (Newell & Simon, 1976)—which bases intelligence on the explicit rule-based manipulation of symbolic

representations—is “fundamentally flawed” (Brooks, 1990, p. 3). Instead, he proposed the “physical grounding hypothesis” (PGH), which stood in direct opposition to mental representations and the computational theory of mind. Brooks argued that “explicit representations and models of the world simply get in the way” (Brooks, 1991b, p. 139); “the world is its own best model... the trick is to sense it appropriately and often enough” (Brooks, 1990, p. 5). Specifically, the PGH states that intelligent systems must have their representations *grounded in the physical world*. Physical grounding, for Brooks, is not simply the linkage between mental representations and the things they signify, but something much more fundamental to the way an agent interacts with its environment. According to this way of thinking, knowledge and understanding is inseparable from environmental interactions and the situational contexts in which they occur<sup>7</sup>.

Brooks’s (1990) approach was also inspired by Minsky, who claimed that intelligence is the product of a collection of simple, hierarchically organized, individually unintelligent, mental agents (Minsky, 1986). Each of these agents (i.e., cognitive processes) is only capable of “some simple thing that needs no mind or thought at all,” (Minsky, 1986, p. 17) but when joined together in “societies” they are capable of producing behaviors that appear intelligent. In accordance with this view, Brooks (1991a) stated, “we hypothesize that *all human behavior* [emphasis added] is simply the external expression of a seething mass of rather independent [behavior generating modules] without any central control or representations of the world” (Brooks, 1991a, p. 226).

---

<sup>7</sup> This viewpoint is often referred to as “situated cognition” (see M. Wilson, 2002, pp. 626–627).

These ideas led to the development of the “subsumption architecture” (see Brooks, 1986) and its implementation in a number of simple, autonomous robots (see Brooks, 1990). The subsumption architecture’s individual “task achieving behaviors” (Brooks, 1986) are implemented using “augmented finite state machines” (AFSMs)<sup>8</sup> that operate asynchronously and without the need for centralized control. These behaviors are connected together in a fixed, layered architecture, where each layer represents a “level of competence” (Brooks, 1986). These competences become progressively more sophisticated as one progresses higher in its vertical stack of layers. The name “subsumption” is based on the fact that each layer can “subsume” (that is, inhibit or suppress) the inputs and outputs of its connected *lower layers* when it wishes to send actuator commands or signals to other layers. In this way, subsumption provides a mechanism for decentralized conflict resolution.

The subsumption architecture side-steps the symbol grounding problem by implementing minds using reactive, stimuli-driven, behavior generating modules. “There are no variables... There are no rules... There are no choices to be made. To a large extent the state of the world determines [an agent’s next] action” (Brooks, 1991b, p. 145). In other words, the subsumption architecture is based exclusively on *online control processes* that are directly coupled to the physical world: they take environmental stimuli as inputs and produce motor commands as outputs, without the use of intervening mental representations or internal models of the world.

---

<sup>8</sup> Brooks (1990) defined an AFSM as a simple finite state machine that is “augmented” with a set of registers (memory buffers for message passing) and a set of timers (that function as alarm clocks).

While the subsumption architecture is capable of producing some seemingly goal-directed behaviors<sup>9</sup>, it is not clear how purely online control processes can implement offline cognitive activities (e.g., deliberation, planning, mental imagery, reminiscing, and dreaming). The direct coupling of environmental stimuli to overt behaviors seems to preclude introspectable mental states or any cognitive process that is spatially or temporally decoupled from immediate sensory stimuli. In many ways, the physical grounding hypothesis and its subsumption architecture embody behaviorist ideals: mental phenomena seemingly vanish under this conceptualization of mind.

Apart from the subsumption architecture, there are other engineering approaches and cognitive theories that could be categorized as non-representational. For example, model-free reinforcement learning (RL) algorithms (e.g., Q-learning; Watkins & Dayan, 1992) employ machine learning techniques to optimize behavioral policies—i.e., functions that map environmental states to actions—without the use of internal environmental models. The intelligent systems produced by model-free RL are consistent with Brooks’s notion of physical grounding.<sup>10</sup> Dynamical systems theory (DST) approaches to cognition, which are based on modeling minds using systems of differential equations, are further examples of non-representational cognitive theories (e.g., see Chemero, 2013; Port & Van Gelder, 1995; Thelen & Smith, 1994).

---

<sup>9</sup> For example, Herbert was an autonomous robot built using the subsumption architecture that could wander through busy office areas, pick up (i.e., steal) soda cans from cluttered desks, and carry them off to be deposited in a trash can near its “home.” (Brooks, 1990 describes Herbert along with many other such robot examples.)

<sup>10</sup> The more recent and capable versions of model-free RL employ function approximators (e.g., neural networks); therefore, they are not strictly “non-representational.” Examples include Deep Q Networks (DQN; Mnih et al., 2015) and Proximal Policy Optimization (PPO; Schulman et al., 2017).



## Connectionism

Connectionist AI (or Connectionism) refers to approaches to building cognitive systems, or explaining mental processes, that are based primarily on *artificial neural networks* (ANNs). ANNs are computational systems inspired by the biological neural networks that occur in brains. And some of the same principles (e.g., distributed representations, parallel processing, and empirical learning) and mechanisms (e.g., activation, activation propagation, and inhibitory/excitatory connections) apply to both artificial and biological neural networks.

ANNs are typically composed of layered sets of computational units called *artificial neurons*. Artificial neurons are highly simplified, mathematically idealized versions of their natural counterparts. Each artificial neuron functions as a computational unit that performs a calculation (e.g., weighted summation) over its inputs (“artificial synapses”). A threshold value (called a bias) is typically added to this intermediate value, and the result is then passed into a non-linear *activation function* to determine an artificial neuron’s output (i.e., its *activation*). These activations may then propagate (spread) over weighted connections (“artificial axons”) to other artificial neurons. While numerous computational elements have been added to this basic design—such as *pooling* (see Boureau et al., 2010), *dropout* (see Hinton et al., 2012), *convolutions* (see Fukushima, 1980; LeCun & Bengio, 1995), and *attention* (Vaswani et al., 2017)—the components described above appear in most, if not all, modern neural network architectures.

In theory, ANNs (with various architectures) have been shown to be universal approximators for many classes of functions (Cybenko, 1988, 1989; Hanin, 2019; Hornik, 1991; Leshno et al., 1993; Lu et al., 2017). However, in practice, there is no single network architecture

that performs well on all tasks. As a result, many special-purpose network architectures have been devised that are optimized for particular input modalities (e.g., visual, auditory, or text) and tasks (e.g., categorization or semantic segmentation). *Convolutional neural networks* (CNNs; Fukushima, 1980; Krizhevsky et al., 2012; LeCun & Bengio, 1995)—which were inspired by Hubel and Wiesel’s work on receptive fields in the visual cortex (Hubel & Wiesel, 1968)—typically work well for visual inputs (e.g., images and videos). While *recurrent neural networks* (e.g., LSTM; Hochreiter & Schmidhuber, 1997) and *transformers* (Vaswani et al., 2017) are often used with sequential (or time series) data (e.g., written language, speech, and music). *Residual networks* (ResNets; He et al., 2016) are optimized for image classification; *U-Nets* (Ronneberger et al., 2015) for the semantic segmentation of medical images; *WaveNets* (Oord et al., 2016) for generating speech and music; and *Generative Pre-trained Transformer 3* (GPT-3; Brown et al., 2020) for text-based natural language processing (NLP) and language comprehension tasks.

Connectionist approaches typically require computationally intensive *training* processes that optimize an ANN for a given task. During training, ANNs are repeatedly exposed to *training examples* (inputs), which are often randomly selected from carefully procured (i.e., cleansed) datasets. An ANN’s parameters (e.g., weights and biases) are then updated based on the network’s performance on those inputs, with respect to a given task. In cognitive science terms, ANNs *learn from experience*, and the goal of these learning processes is to improve an ANN’s overall performance without *overfitting* to their training data. That is, the network’s performance should generalize from a limited set of experiences (i.e., “seen” inputs) to the broader population from which they were sampled (i.e., “never-before-seen” inputs).

ANN training can be broadly classified as either *supervised* or *unsupervised*. Supervised learning uses *labeled* training data. For example, MNIST (see LeCun et al., 1998) is a well-known dataset for supervised learning that pairs images of handwritten digits (inputs) with their corresponding numerical values (labels). Labels (desired outputs) provide a *training signal* that ANNs can use to determine the network’s performance on a given task. Network performance is quantified using a *loss function* (such as, “mean squared error” or “cross entropy”), where *loss*—the loss function’s output—is the basis for updating an ANN’s parameters (i.e., learning). Both a network architecture *and* a loss function are needed to fully specify an ANN’s operation and intent.

Unsupervised learning, on the other hand, uses datasets containing *unlabeled* training examples—that is, only the ANN’s inputs<sup>11</sup>. Consequently, the primary challenge in unsupervised learning approaches is in defining a useful training signal; that is, how does one specify a network’s learning objective in the absence of an explicitly provided, target output value (such as a category label) in the training data? Custom loss functions and network architectures are needed to support this style of learning.

*Autoencoders* are one type of ANN architecture that can learn in an unsupervised fashion. While there are many types of autoencoders<sup>12</sup>—such as, sparse autoencoders (SAEs), denoising autoencoders (DAEs), and variational autoencoders (VAEs)—all of them can be described as the combination of (1) an *encoder* network, (2) a *decoder* network, and (3) a loss function based, in

---

<sup>11</sup> In an autonomous agential system, these inputs might correspond to an agent’s sensory experiences of its environment (e.g., visual, tactile, auditory, gustatory, olfactory, proprioceptive, nociceptive, etc.).

<sup>12</sup> See Bank et al. (2021) for a survey of different types of autoencoders.

part, on *reconstruction error*. The encoder network transforms its inputs into (typically) lower-dimensional representations—called *latent representations*<sup>13</sup> or embeddings. And the decoder network then attempts to *reconstruct* the autoencoder’s original inputs from these latent representations. An autoencoder’s loss function encourages its encoder and decoder networks to work together towards the shared goal of minimizing the differences between the original and reconstructed inputs—that is, to minimize the autoencoder’s reconstruction error. How well autoencoders achieve this objective is a function of (1) how effectively the encoder network learns *generative features* (e.g., shape, size, position, orientation, color) that characterize an autoencoder’s inputs, and (2) how effectively the decoder network can transform those generative features (i.e., latent representations) into likenesses of those original inputs.

Unsupervised learning techniques can be used for feature learning (Bengio et al., 2012), data compression, learning environmental models (Eslami et al., 2018), or training generative processes (among other things). Generative processes are functional components that can render “likenesses” of learned concepts in various modalities (e.g., images, speech, and text).

Variational autoencoders (VAEs; Kingma & Welling, 2013) and generative adversarial networks (GANs; Goodfellow et al., 2014) are two types of generative ANNs that are trained using unsupervised learning.

---

<sup>13</sup> *Latent representations* typically refer to vector encodings of an ANN’s inputs. They are called “latent” because they are usually associated with a network’s hidden layers, rather than its inputs or outputs. For example, the latent representations of autoencoder ANNs are generated in an autoencoder’s (internal) “bottleneck” layer. Latent representations are closely related to “word embeddings”; however, latent representations refer to a broader class of inputs. Latent representations can be thought of as points in a latent (vector) space, where points that are close in this latent space tend to correspond to similar input values.

Having introduced artificial neural networks (ANNs), the question remains: “Can they be used to solve the symbol grounding problem?” The answer to this question is a qualified “Yes.” Internally, ANNs can learn distributed, *non-symbolic representations* that bear a non-arbitrary relationship to their inputs.<sup>14</sup> As a result, ANNs can potentially avoid the infinite regression of symbolic associations—i.e., Harnad’s (1990) “merry-go-round”—that can occur in purely symbolic systems. Moreover, an ANN’s non-symbolic representations can function as *feature detectors* that are sensitive to objects, events, properties, and contextual cues occurring in an environment. The collective *patterns of activation* occurring over these feature detectors can be used to establish concept-referent connections (i.e., grounding connections), which can later be used to identify the concept instances to which those patterns of activation correspond. These representational properties of ANNs can be exploited to implement grounded cognitive processes and systems. (I examine two strategies for doing so later in this chapter.)

Despite these useful properties, if ANNs are trained exclusively on symbolic inputs and outputs (such as words), then their learned internal representations will be *ungrounded*. That is not to say that these representations are “meaningless.” There is often a great deal of information in symbolic contexts (e.g., language) that can be exploited by ANNs. And large language models (LLMs), such as BERT (Devlin et al., 2018) and GPT-3 (Brown et al., 2020), offer compelling evidence that contextual information alone can produce intelligent behaviors. That said, in such cases the network’s knowledge will be “about” those symbols and their uses, not their referents.

---

<sup>14</sup> In Peirce’s terminology, an ANN’s representations can function as icons or indexes. As icons, an ANN’s representations preserve some of the properties of their inputs (i.e., their features). This can be used, for example, to support categorization and categorical inference. As indexes, an ANN’s representations can point to the existence of objects, entities, and events in the environment. This can be used to support predictions about the environment (among other things).

Therefore, the connections between those symbols and their referents will be extrinsic to the system (cf. Searle, 1980). I discuss ungrounded, contextual meaning in more detail later in this chapter (see the section on Distributional Semantics).

ANNs excel at implementing what has been referred to (in dual process theory) as System 1 cognitive processes (Kahneman, 2011; Stanovich & West, 2000). System 1 processes have been characterized as being fast, automatic, heuristic-based, primarily unconscious, and relatively undemanding of computational capacity (Stanovich & West, 2000, pp. 658–659). Perception and intuitive/associative reasoning are System 1 cognitive processes (Kahneman, 2011). On the other hand, ANNs are less adept at implementing System 2 processes, which have been characterized as being slower, analytical, rule-based, primarily conscious, and effortful (Stanovich & West, 2000, pp. 658–659). Algorithmic problem-solving and deliberative reasoning are System 2 cognitive processes (Kahneman, 2011). While, in principle, ANNs should be able to implement both System 1 and System 2 processes, System 2 processes remain more challenging to implement in practice.

ANNs have been criticized for being uninterpretable “black boxes” that require massive amounts of data and computational power to train, and for being slow to adapt to nonstationary environments (see Marcus, 2018 for a summary of these issues). They also suffer from other issues, such as “catastrophic forgetting” (see Goodfellow et al., 2015; McCloskey & Cohen, 1989; Ratcliff, 1990), that frustrate their progress towards general intelligence. As a final note: even if generally intelligent cognitive systems are developed based entirely on ANNs, it is likely that their operations will be largely incomprehensible to us, and they will almost certainly

“think” in distinctly non-human-like ways. Consequently, such intelligent systems may be of limited use to cognitive science.

## Hybrid Systems

Harnad (1990) suggested a possible solution to the symbol grounding problem based on the creation of hybrid symbolic/non-symbolic systems (see Figure 1). Harnad reasoned that symbolic and non-symbolic (e.g., connectionist) approaches have complementary strengths and weaknesses, and that, rather than considering them competing theories of mind, we should combine them. According to Harnad, symbolic systems are incapable of connecting symbols with their referents in the world (i.e., establishing grounding), while connectionist systems lack the *compositionality*, *productivity*, and *systematicity* of symbolic systems (see Harnad, 1990, p. 344). By combining the two, he argued that we not only solve the symbol grounding problem but also retain the benefits of these respective subsystems. One of Harnad’s (1990) central conjectures was that symbols must be grounded *from the bottom-up*, by first learning non-symbolic representations from sensory experiences of objects in the world, and then creating and associating symbolic representations with them.

According to Harnad (1990), symbolic representations must be grounded in two kinds of non-symbolic representations: iconic and categorical.<sup>15</sup> *Iconic representations* are non-symbolic representations that serve as internal “analogs” of the “sensory projections”<sup>16</sup> of objects in the world. These analogs selectively encode many of the distinctive features (e.g., aspects of shape,

---

<sup>15</sup> Harnad’s iconic and categorical representations are both species of *icons* (Peirce, 1867–1893/1992, p. 226) that are distinguished only by their degrees of selectivity and invariance.

<sup>16</sup> The transduction of light in an eye’s retinal cells is an example of a *sensory projection*. Sensory projections occur when sensory stimuli (from an environment) impinge on a sensor.

color, or brightness) of worldly objects, allowing them to be compared with one another based on the similarity of their corresponding iconic representations. *Categorical representations*, on the other hand, are non-symbolic representations that capture the most important *invariant* features corresponding to categories of objects. In particular, they encode the relative weighting of features based on their importance in determining category membership. For example, color may be irrelevant for determining that an object is a chair but highly relevant for determining that a banana is ripe. Harnad (1990) argued that while iconic representations are useful for discriminating between sensory inputs they are insufficient for identifying categories (i.e., types) from category instances (i.e., tokens). In other words, Harnad (1990) contended that categorical representations are needed to support categorical inference.

Harnad (1990) suggested that both iconic and categorical representations should be encoded within the architecture's connectionist (ANN) subsystem. While these representations are useful for identifying and discriminating between sensory stimuli, Harnad (1990) argued that, by themselves, they are an "inert taxonomy" (p. 343) that does not "mean" anything. In Harnad's view, meaning requires that a symbol be assigned to each categorical representation (serving as its "name"), and that these symbols be composed into symbolic *propositions*; for example, symbol strings such as "An X is a Y that is Z" (p. 343). These symbolic representations could then be manipulated based on symbolic rules, as well as the non-symbolic features encoded in iconic and categorical representations.

Harnad's hybrid architecture seems to allow for the best aspects of symbolic and connectionist approaches; however, Harnad's vision leaves open to interpretation many of its details, such as the nature of non-symbolic cognitive operations, how the symbolic and



connectionist components interact, the mechanisms by which iconic and categorical representations influence symbol manipulation, and the functional relationship between iconic and categorical representations (e.g., do categorical representations use or functionally depend on iconic representations in some way?).

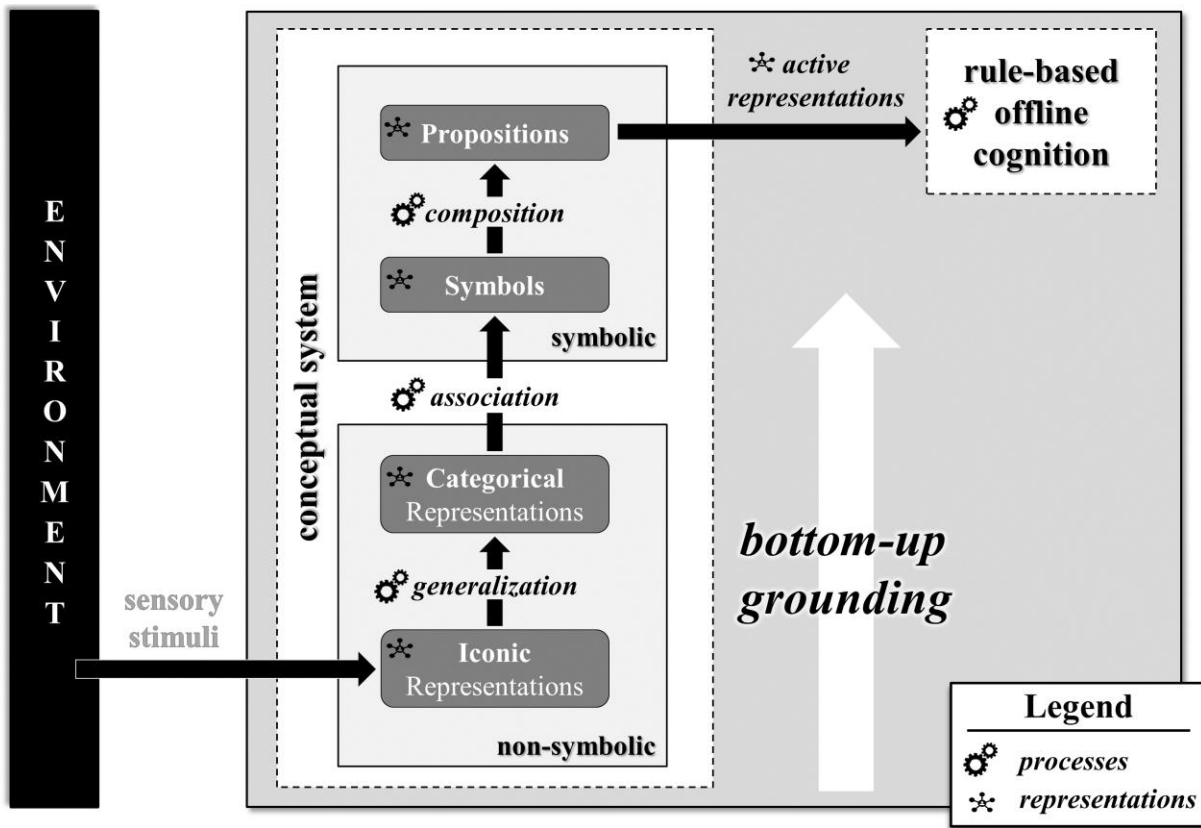


Figure 1. High-level depiction of Harnad's hybrid architecture. Symbolic representations are grounded (bottom-up) in two kinds of non-symbolic representations: iconic and categorical. Each categorical representation is assigned a symbol that serves as its identifier. These symbolic representations are composed into symbolic propositions that can be manipulated by rule-based processes to generate new knowledge.

## Distributional Semantics

Distributional semantics is a family of approaches that bases the meaning of words on the contexts of their use. This idea manifests itself as a cognitive theory called the *distributional hypothesis* (DH; Harris, 1954; Lenci, 2008).

Latent semantic analysis (LSA; Landauer & Dumais, 1997) is perhaps the best known and most influential DH theory in cognitive science. According to the LSA theory, it is the relationships between words that create verbal meaning. Landauer (2007) argued that it is primarily these abstract relations between words that make thinking, reasoning, and interpersonal communication possible. As such, Landauer (2007) believed that LSA was a candidate theory for explaining how verbal meaning is learned, used, and communicated by human minds.

Computationally, LSA describes an algorithm that learns “semantic vectors” for the terms (words) and documents (sequences of words) in a corpus (a set of documents). These vectors can be thought of as specifying locations in a *semantic (vector) space*. And by using *similarity measures* (e.g., cosine similarity), same/difference judgments can be made about these vectors.

Lenci (2008) characterized the semantic vectors learned by LSA (and other DH algorithms) as context-based, distributed, quantitative, and gradual. They are *context-based* because the context of their occurrence determines their semantic interpretation. They are *distributed* because the global distribution of word inter-relations determines their meaning. They are *quantitative* because their semantic interpretation is derived from a statistical analysis over a corpus of language data, rather than a set of qualitative properties. And they are *gradual* because their locations in semantic space gradually update as the global distribution of word contexts in a corpus change.

The mechanics of the LSA algorithm, as described by Landauer and Dumais (1997), are relatively simple. First, a term-document matrix is constructed, where each row corresponds to a term in a corpus and each column corresponds to a document in that corpus. The number of times each term appears in a document determines the matrix's values. Documents are, therefore, encoded as unordered "bags of words," reflecting only the number of occurrences of each word in a document, not their surrounding linguistic context. For example, if the word "car" appeared 17 times in a document, then  $a_{ij} = 17$ ; where  $i$  is the row corresponding to "car,"  $j$  is the column corresponding to that document, and  $A$  is the term-document matrix. The singular value decomposition (SVD; see Strang, 2016) is then calculated for the matrix  $A$ , resulting in the factorization of  $A$  into three matrices:  $A = U\Sigma V^T$ . The matrix  $U$  contains the *semantic (singular) vectors* for terms, the matrix  $V$  contains the *semantic (singular) vectors* for documents, and  $\Sigma$  contains *singular values* that determine the relative weights of the vectors' dimensions. The dimensions that account for the most variance in the data have the largest singular values. Typically, the resulting matrices are truncated by selecting only a subset of each semantic vector's dimensions (e.g., the 100 dimensions with the largest singular values). This *truncated SVD*<sup>17</sup> captures the most important aspects of the corpus's underlying structure while simultaneously removing noise from the dataset.

Landauer and Dumais (1997) were able to show that, given a large enough corpus, LSA can often determine the semantic similarity of words from contextual statistics alone. Thus, they were able to convincingly demonstrate that there is a great deal of information contained within

---

<sup>17</sup> The truncated SVD is also the basis for Principal Component Analysis (PCA), which is widely used in statistics and machine learning for multivariate data analysis and visualization.

the *context* of language usage alone. When thought of as a cognitive model, LSA's ability to infer semantic relationships from contextual statistics raises questions about how humans learn linguistic (and perhaps non-linguistic) concepts. Landauer and Dumais (1997) stated,

the model employs a means of induction—dimension optimization—that greatly amplifies its learning ability, allowing it to correctly infer indirect similarity relations only implicit in the temporal correlations of experience. The model exhibits humanlike generalization... that *does not rely on primitive perceptual or conceptual relations or representations* [emphasis added]. (p. 212)

This idea appears to be in direct opposition to Harnad (1990)'s contention that learning must proceed in a *bottom-up* fashion from sensory and motor experiences. It also casts doubt on the primacy of non-symbolic, perceptual representations in the learning of verbal meaning. Landauer (2007) addressed these points directly, stating

Some assert that meanings are abstract concepts or properties of the world that exist prior to and independently of any language-dependent representations... A sort of corollary of this postulate is that what we commonly think of as the meaning of a word has to be derived from, 'grounded in,' already meaningful primitives in perception or action... Of course, strings of words must somehow be able to represent and convey both veridical and hypothetical information about our inner and outer worlds... However, *once the mappings* [between words and worldly concepts] *have been obtained through the cultural evolution of a language* [emphasis added], there is no necessity that most of the knowledge of meaning cannot be learned from exposure to language itself. (Landauer, 2007, p. 7)

This line of thought led Landauer to conclude that sensory and perceptual experiences might be attached to word meanings *after the fact*, or as he stated it,

This puts the causal situation [i.e., regarding the formation of semantic representations] in a different light. We may often first learn relations of most words and passages to each other from our matrices of verbal experiences and then attach them to perceptual experience by embedding them in the abstract word space. (Landauer, 2007, p. 7)

In Landauer's view, this abstract word space (i.e., semantic space) is the "primitive substrate" that gives words meaning. He goes on to suggest that semantic spaces could also be constructed for percepts, and that, after "only a comparatively few correlations" between perceptual and linguistic experiences, all of the connections between these semantic spaces will be aligned "to a close approximation" (Landauer, 2007, p. 24).

To sum up, the distributional hypothesis (DH) is a cognitive theory based on the idea that contextual relationships are the primary mechanisms by which we acquire verbal meaning (and perhaps other types of meaning; see Landauer, 2007). And latent semantic analysis (LSA) is one of the best-known and influential of these theories. As a computational approach, LSA has conclusively demonstrated that the co-occurrence patterns of words in a corpus can be used to infer the semantic similarities between those words given a large enough structured corpus. As a cognitive theory, LSA predicts that concept and word acquisition is largely a function of learning the semantic relations between words through repeated exposure to language. In this view, language is seen as already containing most of the necessary information for learning these concepts. Landauer (2007) hypothesized that perceptual experiences could then be associated with these linguistic forms, once the underlying semantic spaces have been learned.

I contend that DH-based algorithms, such as LSA, learn a form of “associative meaning” that captures the inter-connectedness of things, not the things themselves. In particular, I believe that these approaches demonstrate that it is possible to learn *how* and *when* to use words without knowing *what* they refer to in the world. A similar phenomenon has been seen with recent massive-scale<sup>18</sup>, neural network-based, language models—like *Generative Pre-trained Transformer 3* (GPT-3; Brown et al., 2020)—that have demonstrated human-level performance on various language comprehension tasks without the use of grounded representations.

While distributional semantics will likely have some part—perhaps a large part—to play in explaining how humans learn language, it does not solve the symbol grounding problem. And, in my opinion, it does not obviate the need to do so. In particular, LSA and distributional semantics say very little about the nature of non-linguistic thought (e.g., non-symbolic perception and action, and cognitive processes like mental simulation). Consequently, distributional semantics, by itself, does not explain the thought processes of non-human animals and pre-verbal humans. That said, the combination of distributional semantics with grounded approaches to cognition is a promising research direction worth exploring. And numerous researchers have advanced theories of cognition that suggest that humans may use a combination of grounded representations and distributional semantics (e.g., Barsalou et al., 2008; Landauer, 2007; Louwrese, 2018; Louwrese & Jeuniaux, 2008).

---

<sup>18</sup> GPT-3 has 175 billion model parameters, and was trained on half-a-trillion encoded linguistic tokens. It received a great deal of media attention because it has been claimed that the *synthetic* news articles generated by the model are practically indistinguishable from *real* news articles. For example, Brown et al. (2020) noted, “mean human accuracy at detecting articles that were produced by the [GPT-3] 175B parameter model was barely above chance.”

## **Embodied, Simulation-Based Cognition**

Embodied cognition (EC) refers to a set of related though distinct ideas that emphasizes the fundamental role of bodies and environmental interactions in shaping minds. Rather than conceptualizing minds as abstract information processing machines that are distinct from bodies, EC theories contend that mental activities are inseparable from the sensory and motor mechanisms of the body. In general, many of the fundamental characteristics of EC can be seen as emerging in response to, and in opposition of, the computational origins of cognitive science.

Margaret Wilson (2002) identified six distinct perspectives that characterize embodied cognition:

- (1) “Cognition is Situated,”
- (2) “Cognition is Time Pressured,”
- (3) “We Off-Load Cognitive Work onto the Environment,”
- (4) “The Environment is Part of the Cognitive System,”
- (5) “Cognition is for Action,” and
- (6) “Off-Line Cognition is Body Based.”

I will focus here on the first and sixth of these perspectives, as they motivate the development of embodied, simulation-based cognition. However, the fifth of these perspectives (“Cognition is for Action”) is also of critical importance to this work, and it appears both explicitly and implicitly throughout.

“Cognition is Situated,” refers to the idea that cognitive activity takes place with respect to the task-specific and context-sensitive demands of an environment. Situated agents are said to be tightly *coupled* to their environments, and their cognitive processes are continually influenced by incoming sensory stimuli and the effects of their actions on their environments. Proponents of situated cognition tend towards a view of the *action-perception cycle* (Cutsuridis et al., 2011; T. M. H. Dijkstra et al., 1994; Fuster, 2004) that is rapid and reactive, minimizing or eliminating many notions of *offline* cognition, such as inference, planning, and deliberative thought. Sensory stimuli (inputs) are more or less directly connected to corresponding behavior producing modules (outputs) that allow agents to act on their environments. “Any cognitive activity that takes place ‘off-line,’ in the absence of task-relevant input[s] and output[s], is by definition not situated” (M. Wilson, 2002, p. 626). By contrast, situated activity occurs *online*. Computationally, situated cognition is well-exemplified by Rodney Brooks’s subsumption architecture (see the section on Non-Representationalism presented earlier in this chapter).

While situated cognition is an important perspective within embodied cognition, it fails to provide a convincing explanation for the many cognitive activities that appear to occur offline, decoupled from immediate sensory inputs and motor activity. Offline cognitive activities, such as planning, deliberation, prediction, inference, imagination, and even more pedestrian activities like the recall of long-term memories, are *decoupled* from an agent’s immediate inputs and outputs that are occurring “right now.” Even though these cognitive activities are commonplace, and seemingly indispensable for humans (and likely other animals), much of the embodied literature has chosen to deny them, ignore them, or downplay their importance. This is likely due, at least in part, to the antagonistic relationship that exists between embodied views and



more traditional cognitive theories (e.g., the CTM and the LOTH) that have focused almost exclusively on offline activities using ungrounded, symbolic representations. However, the difficulty of explaining these processes in a way that is grounded and consistent with embodied principles is almost certainly a contributing factor to their relative neglect in the embodied cognition literature. Margaret Wilson's sixth perspective, "Off-Line Cognition Is Body Based," corresponds to an effort within embodied cognition to explain how agents' bodies and environmental interactions can be integral to these offline cognitive processes.

Researchers in cognitive science, linguistics, neuroscience, and philosophy (including Lawrence Barsalou, Benjamin Bergen, Vittorio Gallese, Arthur Glenberg, Alvin Goldman, Germund Hesslow, Marc Jeannerod, Jesse Prinz, and Giacomo Rizzolatti) have advanced various theories about how embodiment can be extended to offline cognition through the use of some form of mental simulation. I generically label these efforts with the term *embodied, simulation-based cognition*, which is the central topic of this manuscript. According to this view, sensorimotor mental simulations, and the processes that operate on them, can be viewed as mechanisms for realizing "grounded cognition" (Barsalou, 2008) in natural and artificial minds. Margaret Wilson (2002) summarized this idea, stating that the

[m]ental structures that originally evolved for perception or action appear to be co-opted and run 'off-line,' decoupled from the physical inputs and outputs that were their original purpose, to assist in thinking and knowing.... In general, the function of these sensorimotor resources is to run a simulation of some aspect of the physical world, as a means of representing information or drawing inferences. (M. Wilson, 2002, p. 633)

In other words, these theories propose a view of cognition based on grounded, non-symbolic representations, mental simulations, and imagistic processes (e.g., image generation, transformation, and introspection; see Chapter 7).

Many theories have been advanced within this tradition. Some focus on specific cognitive functions, such as language comprehension (Bergen & Chang, 2005; Bergen & Wheeler, 2010; Zwaan, 2004; Zwaan & Taylor, 2006), action and motor cognition (Jeannerod, 1995, 2006; Jeannerod & Frak, 1999), and “theory of mind” (Gallese, 2003, 2007; Iacoboni et al., 2005; Rizzolatti & Sinigaglia, 2016; Shanton & Goldman, 2010). Other theories are more comprehensive, proposing that mental simulation underpins most, if not all, cognitive functions (Barsalou, 1999, 2008, 2009; Grush, 2004; Hesslow, 2002, 2012; Prinz, 2004). I will primarily focus on Barsalou’s (1999) theory of perceptual symbol systems and Jeannerod’s (1994, 1995, 2001) theory of motor cognition here, as they heavily influenced the theory and computational implementations developed in later chapters.

### ***Perceptual Symbol Systems***

According to the theory of perceptual symbol systems (PSSs), the patterns of activation that occur in sensory and motor systems during perception and action can be stored into long-term memory and, if later recalled, function as grounded representations of entities and concepts in the physical world. Barsalou referred to these re-instantiations of sensorimotor patterns of activation as *perceptual symbols* (or modal symbols). Despite their name, perceptual symbols are actually *non-symbolic* representations.

Barsalou (1999) outlined numerous properties that he believed would characterize perceptual symbols.<sup>19</sup> Perceptual symbols are *modal*, because “they are represented in the same systems as the perceptual states that produced them,” (Barsalou, 1999, p. 578) and, potentially, *multimodal*, accounting for content originating in multiple sensory modalities. They are *analogical* because they share properties with, and likely bear some structural resemblance to, their originating perceptual states. In contrast, *amodal* symbols typically have an arbitrary relationship with the phenomena to which they refer.

Perceptual symbols are *not complete “recordings”* of an individual’s entire mental state during perception, but instead capture salient aspects of their corresponding sensory and motor experiences. Barsalou characterizes the information represented in perceptual symbols as “relatively qualitative and functional (e.g., the presence or absence of edges, vertices, colors, spatial relations, movements, pain, heat)” (Barsalou, 1999, p. 582), and he is careful to point out that perceptual symbols should not be thought of as physical “pictures.” Further consequences of this are that perceptual symbols may contain aspects that are *partial and indeterminate*; for example, the perceptual symbol for a bicycle wheel may have an unspecified number of spokes.

The same perceptual symbol can be used to *designate multiple referents* since resemblance, by itself, is not enough to establish intentionality (i.e., the aboutness of a mental representation). Rather, mechanisms external to perceptual symbols must be used to establish their relationships to specific referents (e.g., contextual information). To illustrate this, consider attending a dinner party where all of the guests have wine glasses of identical design. In such a

---

<sup>19</sup> The Appendix contains a much more detailed review and analysis of various representational properties ascribed to perceptual symbols.

scenario, the designation of “my wine glass” is established both by its appearance and its context (for example, its placement on the table), not resemblance alone.

Perceptual symbols are *dynamic*; that is, they are sensitive to differences in external and internal (i.e., mental) contexts, as well as changes in nearby regions of associative memory. As a result, the reactivation of a perceptual symbol has dynamical properties that result in *a degree of variability in their reconstructions* from long-term memory. Barsalou stated that perceptual symbols can be viewed as “attractors” (see Norton, 1995, p. 56) in a connectionist network: “As the network changes...the attractor changes. As the context varies, activation of the attractor covaries” (Barsalou, 1999, p. 584).

Barsalou (1999) equated the capacity to mentally simulate a concept with an understanding of that concept: “Once individuals can simulate a kind of thing to a culturally acceptable degree, they have an adequate understanding of it” (Barsalou, 1999, p. 587). In particular, he argued that once an agent has enough conscious experience with some entity, object, or event, the set of perceptual symbols corresponding to that concept can become integrated into a *simulator*. These simulators constitute “the knowledge and accompanying processes that allow an individual to represent some kind of entity or event adequately” (Barsalou, 1999, p. 587), and their purpose is to produce *mental simulations* that are non-symbolic instantiations of some conceptual “type” (e.g., category). These sensorimotor

simulations are typically *preconscious*<sup>20</sup> representations that occasionally become conscious. When they do become conscious, they are referred to as *mental images*.

Just as “a pile of bird features does not make a bird” (Murphy, 2004), a set of perceptual symbols does not define a concept. To solve this problem, Barsalou suggested that simulators must structure their perceptual symbols within (background or reference) *frames*<sup>21</sup>. The role of such a frame is to arrange a simulator’s perceptual symbols in “just the right way” so that their relationships to one another (in the context of a particular concept) are adequately specified and constrained. For example, a frame for the concept of a stationary bicycle might include the spatial relationships between each of its subcomponents (for example, the bicycle’s wheels, frame, seat, and handlebars) in relation to its overall volumetric extent. To represent dynamic concepts, such as a bicycle in motion or a song, frames may also need to include a temporal dimension. Barsalou’s frames have some functional overlap with what Kosslyn et al. (2006) called *object maps* (or spatial images); however, Barsalou’s frames include additional properties, such as “predicates” and “constraints,” that extend beyond Kosslyn et al. (2006)’s relatively simple conceptualization of object maps.

---

<sup>20</sup> I use the convention established by Franklin and Baars (2010) of referring to unconscious representations that have the *potential* to become conscious as “preconscious” and those that do not as “never-conscious.”

<sup>21</sup> Barsalou’s frames are similar in spirit, but different in kind, to Minsky’s (1975) frames. Where Minsky’s frames are amodal, symbolic representations, Barsalou envisions frames as modal, non-symbolic representations.

Barsalou (1999) was adamantly opposed to amodal symbols<sup>22</sup>, even if they could be grounded as Harnad (1990) proposed using a hybrid symbolic/non-symbolic architecture.

Barsalou criticized Harnad's approach saying,

One solution [to the symbol grounding problem] is to postulate mediating perceptual representations. According to this account, every amodal symbol is associated with corresponding perceptual states in long-term memory... During symbol grounding, the activation of the amodal symbol in turn activates associated perceptual memories, which ground comprehension. *Problematically* [emphasis added], though, perceptual memories are doing all of the work, and the amodal symbols are redundant. *Why couldn't the system simply use its perceptual representations ..., both during categorization and reasoning* [emphasis added]?

Consequently, Barsalou (1999) has argued that modal representations and mental simulations are sufficient to implement a “fully functional conceptual system” without the need for amodal (i.e., ungrounded) symbols, such as those used in classical symbolic AI and CTM. Specifically, Barsalou (1999) attempted to demonstrate that a PSS could support types and tokens, categorical inference, productivity, propositions, and abstract concepts.

Barsalou (1999) further criticized amodal approaches saying that they (1) “failed to provide a satisfactory account of the transduction process that maps perceptual states into amodal

---

<sup>22</sup> Goldstone and Barsalou (1998) characterized perceptual symbol systems as an “eliminativist position” on the uniting of perception, action, and cognition, that attempts to establish an “existence proof that a completely perceptual approach is sufficient for establishing a fully functional symbolic system” (Goldstone & Barsalou, 1998, p. 235). They argued that, if this position is determined to be well-founded, it is grounds for eliminating amodal symbols as unnecessary theoretical baggage.

symbols” (p. 580); (2) “have not fared well in implementing... spatio-temporal knowledge” (p. 580); (3) are not supported by neuroscientific and cognitive evidence; and (4) can only provide *post hoc* explanations of mental phenomena (such as timing effects in mental rotation and scanning) that modal theories predict a priori.

Moreover, Barsalou contended that linguistic symbols are *not* amodal symbols, “nor does an amodal symbol *ever* develop in conjunction with [them]” (Barsalou, 1999, p. 592). Instead, he viewed linguistic symbols as *modal*, corresponding to sensory experiences of word forms (e.g., how they look and sound). These become associated with other simulators corresponding to their referents (i.e., what these linguistic elements correspond to in the world), and it is this combination of word form and referent simulators that allow “linguistic control over the construction of simulations” (Barsalou, 1999, p. 582). Finally, he stated that while it is often argued that amodal symbols acquire meaning by associations with other amodal symbols (cf. Distributional Semantics), this approach ultimately fails without grounding, terminal symbols.

Conceptually, the theory of PSSs makes several important theoretical innovations:

- (1) It suggests an operational definition of (modal) understanding based on an individual’s capacity and facility at simulating concepts.
- (2) It provides a mechanism by which new knowledge can be produced *offline* using modal representations, mental simulation, and imagistic thought processes.
- (3) It pushes the boundary on what can be accomplished with modal, analogical representations, suggesting that non-symbolic, imagistic thought could be the backbone of natural, and perhaps one day artificial, intelligent systems.

Unfortunately, to date, the theory of PSSs has failed to compellingly demonstrate that modal, non-symbolic representations alone can account for *all* cognitive phenomena (e.g., abstract concepts; see Dove, 2009). As such, the onus still falls on Barsalou to demonstrate the redundancy of amodal symbols. Furthermore, like Harnad's (1990) hybrid architecture, PSS is a high-level theory that is intuitively appealing but underspecified, and it requires more computational accounts to determine the feasibility of implementing this theory in practice (Barsalou, 2009; Pezzulo et al., 2013).

### ***Motor Cognition***

Jeannerod (1994, 1995, 2001) advanced a simulation-based theory of cognition that attempts to unify the production of actions and action-oriented thought. According to this theory, action production involves both an overt stage, in which actions are executed, and a covert stage, in which actions are mentally simulated. Jeannerod claimed that covert actions (motor simulations) are the cognitive precursors of overt actions, and that motoric mental imagery occurs through the enactment of covert actions without a subsequent overt action execution stage (Jeannerod, 2001, p. S103)<sup>23</sup>. Jeannerod hypothesized that the covert stages of action production generate mental representations that include an action's goal (i.e., its intention), a means to reach that goal (i.e., an action plan), and the action's environmental consequences.

There is a wealth of psychological and neurophysiological evidence suggesting that *thinking about acting* (i.e., imaging oneself performing external activities) and *actually acting*

---

<sup>23</sup> This raises the question of the mechanism by which action execution is prevented during motoric mental imagery. Recent evidence suggests that biological systems have neural circuits that directly inhibit motor activity during motor simulation, thus actively preventing overt action execution (Angelini et al., 2015; Rieger et al., 2017; Scheil & Liefoghe, 2018). An alternate hypothesis, which argued that covert actions lack sufficient activation to engage motor execution, has been largely discredited.



(i.e., performing external activities) engage similar cognitive mechanisms and utilize shared neural pathways. For example, when subjects were asked to imagine themselves performing an action (e.g., walking), they typically required an amount of time that is proportional to the time required to physically perform that action (Bakker et al., 2007; Decety et al., 1989; Frak et al., 2001). Furthermore, neuroimaging studies—based on fMRI, PET, and near-infrared spectroscopic (NIRS) topography—have demonstrated that similar patterns of neural activity occur in subjects' primary and pre-motor cortices when imagining and performed activities such as moving one's fingers, toes, and tongue (Decety et al., 1994; Ehrsson et al., 2003; Lotze et al., 1999; Porro et al., 1996; Sharma & Baron, 2013). Taken together, these results suggest that motoric mental imagery and physical activity operate using similar cognitive processes and a shared neural substrate.

Building on these findings, researchers have noted that simply *observing* the actions of others can elicit a pattern of neural activity that resembles the mental simulation of one's own actions (Buccino et al., 2013; Calvo-Merino et al., 2005, 2006; Gallese et al., 1996; Rizzolatti et al., 1996; Rizzolatti & Sinigaglia, 2016). The neurons that perform this dual function are often referred to as “mirror neurons” (Rizzolatti & Craighero, 2004) and their associated cognitive processes as “mirror mechanisms” (Rizzolatti & Sinigaglia, 2016). Rizzolatti and Sinigaglia stated,

The functional properties of the mirror mechanism indicate that the motor processes and representations that are primarily involved in generating and controlling a given behaviour can also be recruited in an individual who is observing someone else displaying that behaviour. By means of this recruitment, the individual may take

advantage of his or her own processes and representations to understand others' actions and emotions, as well as their corresponding vitality forms. (Rizzolatti & Sinigaglia, 2016, p. 7)

It has been shown that these “mirror mechanisms” can also be engaged when trying to understand the *intentions* of others (Iacoboni et al., 2005), and during the comprehension of motoric, action-oriented language (Jirak et al., 2010; Zwaan & Taylor, 2006).

## Chapter 3

### A Hybrid Approach to Embodied, Simulation-Based Cognition

Symbols grow. They come into being by development out of other signs, particularly from likenesses or from mixed signs partaking of the nature of likenesses and symbols....

These mental signs are of mixed nature. (Peirce, 1893–1913/1998, p. 10)

This chapter is an overview of **ES-Hybrid (Embodied Simulation-Hybrid)**—a *hybrid* (symbolic/non-symbolic) approach to embodied, simulation-based cognition (ES). It is intended to serve as a roadmap for contextualizing the work in subsequent chapters.

A fundamental assumption shared by all ES theories is that individuals conceptualize, understand, and act upon their environments using representations and processes that are grounded in their sensory and motor systems. That is, the mental representations and cognitive processes used to interact with the world are recruited to reason about the world. Consequently, offline cognition in ES largely entails the re-enactment and transformation of prior sensorimotor states and their representational derivatives. Given this, all ES theories should *minimally* specify (1) how knowledge can be grounded in sensory and motor systems; (2) how knowledge can be acquired, both empirically and through reason; (3) how knowledge can be brought to bear in support of simulation-based offline cognition; and (4) how simulation-based cognition can support the production (i.e., the selection and execution) of actions. Developing ES-Hybrid's perspective on these basic questions is the focus of this manuscript.

In this chapter, I discuss the circumstances that led to and motivated ES-Hybrid's creation, acknowledge the theories and ideas that influenced its development, and discuss some

of its limitations which will be addressed in subsequent chapters. Following this, I provide a high-level overview of the theory—a narrative that introduces ES-Hybrid’s mental representations and cognitive processes and describes how they collaborate to realize a hybrid conceptualization of ES. Many of ES-Hybrid’s components are then individually discussed and elaborated on. I conclude the chapter with brief discussions on intentionality, intrinsic meaning, and language in the context of ES-Hybrid.

## **Preliminaries**

### *Origins and Motivations*

ES-Hybrid developed from an initial attempt to implement Barsalou’s theory of perceptual symbol systems within the LIDA cognitive architecture (e.g., see Kugele & Franklin, 2020b). As a result of that exercise, I came to regard Barsalou’s theory as both deeply inspired and conspicuously flawed. The fundamental principles of embodied, simulation-based cognition that developed out of that theory are intuitively appealing and well-supported empirically (see Barsalou, 2008, pp. 623–631). Nevertheless, Barsalou’s “eliminative position” (see Goldstone & Barsalou, 1998) regarding amodal and ungrounded representations seems largely indefensible.<sup>1</sup>

As Toomela (1999) observed, there is knowledge about the world that is qualitatively out of the reach of our senses. We will never directly observe the “strings” from string theory, gravitational singularities, transfinite numbers, or consciousness. These concepts are

---

<sup>1</sup> A strict implementation of a perceptual symbol system is also largely incompatible with LIDA, which is a hybrid (symbolic/non-symbolic) cognitive architecture. Moreover, Stan Franklin (the creator of LIDA) thought it likely that some concepts (e.g., transfinite numbers) are ungrounded, though he regarded them as far less common than their grounded counterparts (S. Franklin, personal communications, 2020-2022).

permanently ungrounded by the nature of their content. Nevertheless, this does not imply that such things are beyond the reach of our thoughts.

Moreover, we can think about objects, entities, and events for which grounding is possible *in principle*, but it has yet to be established empirically within one's own mind. These are hypotheticals (i.e., predicted objects, entities, situations, and events). They have yet to be directly observed, therefore, they lack grounding content. And yet we are capable of reasoning about them as well.

Consider, for example, the mental representations for events that have *unknown causes* (e.g., idiopathic diseases), words that have *unknown meanings* (e.g., “floccinaucinihilipilification”), and objects that have *no known properties* (e.g., “dark matter”). The existences of these *unknown referents* can be inferred—in the absence of any direct observations—from their indexical relationships with other things. Diseases have causes, words have meanings, and objects have properties; therefore, we assume those same principles must also apply in these instances. While these inferred mental representations *initially* lack specific identities and tangible properties, we can, nevertheless, think about such things. Therefore, distinct mental representations must exist in our minds for these referents.

According to the theory of perceptual symbol systems (PSSs), all mental representations are modal, analogical, and grounded (Barsalou, 1999). But what sensorimotor experiences can be constitutive of these unknown referents? What properties or structures exist in their corresponding perceptual states that could be used to establish an analogical (resemblance-based) relationship with those referents? How might these representations be grounded at the time of their creation when their existences were inferred, not observed? Contrary to the theory of PSSs,

I contend that such representations are non-analogical, amodal, and *initially ungrounded*. In other words, they are symbolic (in the Peircian sense; see Chapter 2).

Furthermore, I contend that these initially ungrounded, amodal symbols are not fringe phenomena. They are the pervasive byproducts of top-down, predictive processing. Amodal symbols are created (or otherwise allocated) by a cognitive system whenever the existence of an object, entity, or event is inferred.<sup>2</sup> These amodal representations can then be used as “cognitive placeholders” within that system’s associative machinery.

Such ungrounded symbols are not functionally inert. Their very existence can influence one’s actions, demanding answers to the epistemic questions they pose—e.g., “what does floccinaucinihilipilification mean?” More specifically, *ungrounded representations* can exert pressure on a cognitive system that encourages the selection of “epistemic actions” (see Kirsh & Maglio, 1994) that are intended to ground them. And, in most cases, these amodal, symbolic representations can *eventually* be grounded via active exploration and speculative reasoning.

ES-Hybrid’s approach to abstract concepts is a simple extension of these ideas. The only difference being that the representations for abstract concepts are *ungroundable*. It is not simply a matter of acquiring the right sensorimotor experiences—grounding experiences do not exist for these concepts. Therefore, their meanings must be established through non-grounding associations.

---

<sup>2</sup> I also suspect that amodal representations may more generally support multimodal binding (similar to Damasio’s convergence zones or hub-and-spoke representational models; see Damasio, 1989; Patterson et al., 2007; Ralph et al., 2010, 2017).

I contend that amodal (symbolic) representations are ubiquitous in concept representation. However—perhaps counter-intuitively—I also contend that they are *always* associated with supporting modal (non-symbolic) representations. This necessitates a conceptual shift. Rather than thinking about concepts in terms of *atomic* representations, we need to think in terms of *composite* representations. Rather than asking whether a concept’s representation is symbolic or non-symbolic, we should assume that *all* concept representations contain elements of both.<sup>3</sup>

Using combinations of modal and amodal representations, ES-Hybrid can represent concrete and abstract concepts, support unknown referents, perform multimodal sensorimotor bindings, and implement contextual disambiguation (i.e., solve the designation problem; see Barsalou, 1999, sec. 2.2.3). Furthermore, these hybrid representations naturally support “active perception” and “cognitive indirection.”<sup>4</sup> Finally, ES-Hybrid’s amodal representations can serve as “scaffolding” for its modal representations, and vice versa.

As a further motivation for ES-Hybrid, the action-oriented aspects of cognition are generally glossed over by Barsalou. According to the tenets of embodied cognition, cognition is ultimately in service of action (Franklin, 1995, Chapter 16; M. Wilson, 2002, pp. 631–632); therefore, this aspect of cognition should be addressed. Moreover, many offline cognitive

---

<sup>3</sup> Michel (2021) made a similar case when he argued that the modal/amodal dichotomy should be reconceptualized as a *graded* representational property. (In particular, see Michel, 2021, sec. 6.2 for a good discussion of this idea.)

<sup>4</sup> I define *cognitive indirection* as the ability to rapidly adjust a mental representation’s referential and non-referential associations, while maintaining any dependent associations (i.e., the other representations that refer to it). In this way, a representation’s properties, grounding, and intentionality can be fluidly changed while simultaneously maintaining other previously established associative relationships. Cognitive indirection is useful for considering alternate explanations for unknown referents, among other things. (I discussed this again in the context of an example later in this chapter.)

processes can be modeled as the simulated execution of internalized overt behaviors and their environmental consequences (Gallese et al., 1996; Jeannerod, 2001; Rizzolatti et al., 1987). In this sense, cognition *is* action, and thinking can be modeled as a skill (Bartlett, 1958).

### ***Influences***

ES-Hybrid incorporates ideas from Barsalou's theory of perceptual symbol systems (Barsalou, 1999), Harnad's (1990) hybrid grounded architecture, and Jeannerod's (2001) theory of motor cognition. Lakoff and Johnson's (1999, 1980/2008) embodied theory of analogical thought inspired portions of ES-Hybrid, and situated, online control processes (Brooks, 1986, 1990, 1991b) also figure into this theory. (Many of these ideas were covered in Chapter 2.)

Finally, it is hard to overstate LIDA's (Franklin et al., 2016) influence on ES-Hybrid. ES-Hybrid was conceived of and developed with respect to an eventual implementation within the LIDA cognitive architecture. As a result, ES-Hybrid can be viewed as an abstraction over LIDA, and it naturally reflects LIDA's conceptual commitments and its assumptions about the nature of cognition.

### ***Limitations***

ES-Hybrid is as a *high-level* hybrid account of embodied, simulation-based cognition (ES). It is intended to serve as a "meta-architecture" that informs the modeling and implementation of specific aspects of ES. As such, it leaves many details *unspecified* (for example, attention and learning mechanisms) and others *underspecified* (for example, motivational systems and action selection). It makes no strong commitments to any modular organization of mind (e.g., short- and long-term memory modules, or perceptual systems). Nor does it mandate specific



implementations for many of its cognitive processes. Consequently, ES-Hybrid is too abstract to serve as a “unified theory of cognition” (see Newell, 1994).

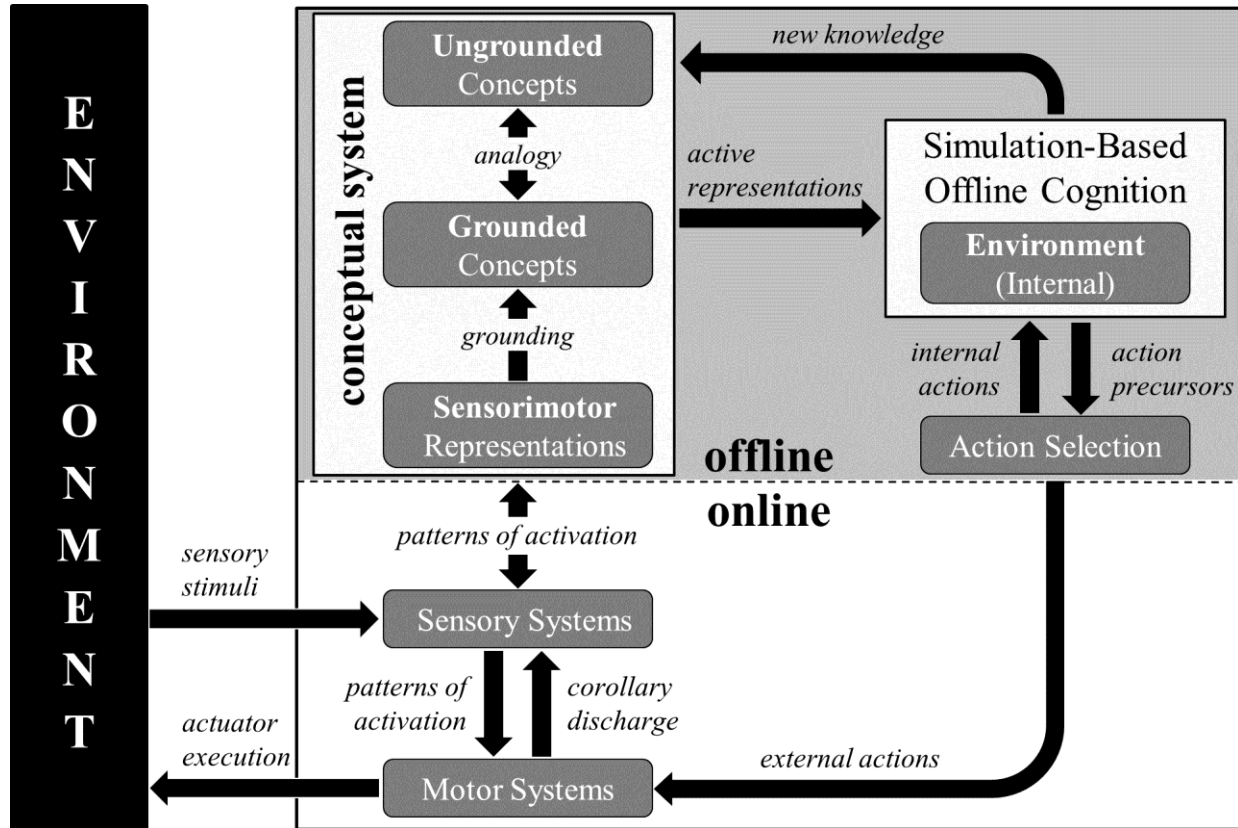


Figure 2. Illustration of ES-Hybrid’s major components and their interactions.

This partly reflects the fact that ES-Hybrid was always intended to be implemented within a cognitive architecture, specifically LIDA.<sup>5</sup> However, this also reflects a desire to avoid overly constraining its implementations through unnecessary design choices. In general, ES-Hybrid is intended to be a minimalistic system specification in order to maximize the flexibility and extensibility of its future applications.

<sup>5</sup> The primary focus of the remainder of this manuscript will be the conceptual and computational implementations of fundamental ES-Hybrid components within the LIDA cognitive architecture.

## Overview

Figure 2 shows the major components of ES-Hybrid, including its basic mental representations, cognitive processes, and their high-level relationships to one another. Many of these will be described in the text below.

*Sensory stimuli*, originating from an agent’s (external) environment, are transduced by an agent’s sensors, activating modality-specific feature detectors. The resulting *patterns of activation* are encoded as *sensorimotor representations* that can serve as “sensory signatures” for those environmental stimuli. Resemblance-based *generalization processes* can then operate on these sensorimotor representations to construct increasingly invariant *grounded representations*.

Grounded representations signify sense-able objects, entities, and situations in an agent’s (external or internal) environment. These representations can vary in their perceptual/conceptual specificity, from viewpoint-specific, context-sensitive, sensory experiences (e.g., a sunset viewed in a particular moment from the window of a plane) to more general *concrete concepts* (e.g., sunsets at large). However, they are *always* grounded in sets of modality-specific, sensorimotor representations.

*Ungrounded representations* are created when *predictive processes* infer the existence of “unknown referents”—hypothesized objects, entities, situations, and events that have yet to be directly observed. While these representations are *initially ungrounded*, they are always associated with supporting modal content (via non-referential associations, which are described later). This modal content serves an indexical relationship with those ungrounded amodal components, pointing to their existence rather than depicting their content. In most cases, these ungrounded structures are *eventually grounded* through direct observation or speculative

reasoning. However, in a minority of cases, a structure may be inherently *ungroundable*; that is, the things to which these representations refer can be inferred but not directly sensed in an environment. This is how ES-Hybrid models *abstract concepts*.

An important aspect of ES-Hybrid is its distinction between referential and non-referential associations. *Referential associations* are grounding associations that establish links between representations and the things to which they refer. *Non-referential associations* are non-grounding associations that represent non-correspondence-based relationships such as causality, co-occurrence, and sequential ordering. Non-referential associations between concrete and abstract concept representations can enable *analogical* (e.g., metaphorical) reasoning.

*Simulation-based, offline cognitive processes* manipulate *active representations*<sup>6</sup> for the purpose of generating *new knowledge* and the representational precursors of action (e.g., beliefs, desires, intentions, and situational contexts). A fundamental assumption of this theory is that offline cognition is primarily based on imagistic, epistemic (knowledge-generating) processes and the execution of internal (covert) actions (Jeannerod, 2001). These cognitive processes are not based on explicit, rule-based, symbolic computations, but on the generation, transformation, and inspection (see Kosslyn, 1994; Kosslyn et al., 2006) of internally perceivable, sensory-like content (i.e., mental simulations).

An *action selection* mechanism selects external or internal actions based on the *action precursors* resulting from offline cognition. Jeannerod (2001) referred to externally directed

---

<sup>6</sup> Active representations are situationally relevant representations that are “activated”—i.e., made available or more salient to cognitive processes—based on the situational elements embedded within an agent’s internal and external environment. While I generally assume that perception, memory recall, and mental simulation contribute to these activations, the details are left unspecified here.

actions as “overt” actions and internally directed actions as “covert” actions. Covert actions generate action-based mental simulations that can be used to predict the consequences of an agent’s actions. They can also be used to understand the actions and intentions of others (see Gallese et al., 1996; Iacoboni et al., 2005; Rizzolatti et al., 1996). More generally, this capability to execute covert actions is the basis for “motor cognition” (see Jeannerod, 2001, 2006) and the volitional aspects of mental imagery (see Chapter 7). Covert actions can be thought of as effecting changes in an agent’s *internal environment*—a transient, internal model of an agent’s current situation.<sup>7</sup>

The execution of overt (external) actions occurs through a process of *situated, online control* (cf. Brooks, 1990, 1991b). These online cognitive processes directly receive the patterns of activation induced in sensory systems by environmental stimuli, and they send appropriate low-level motor commands to an agent’s actuators (e.g., exciting muscle fibers). That is, they are directly coupled to an agent’s environment through its sensory and motor systems. Once an action is selected, online cognitive processes continue to fulfill the execution of that action until changes in an agent’s internal state result in the selection of a different action. In general, *offline cognition* can be seen as orienting or disposing a cognitive system towards fulfilling a proximal intention, while *online cognition* fulfills that proximal intention through overt behaviors.

---

<sup>7</sup> While an agent’s internal environment can be spatially and temporally decoupled from its external environment, it often mirrors and augments the recently observed portions of an agent’s external environment.

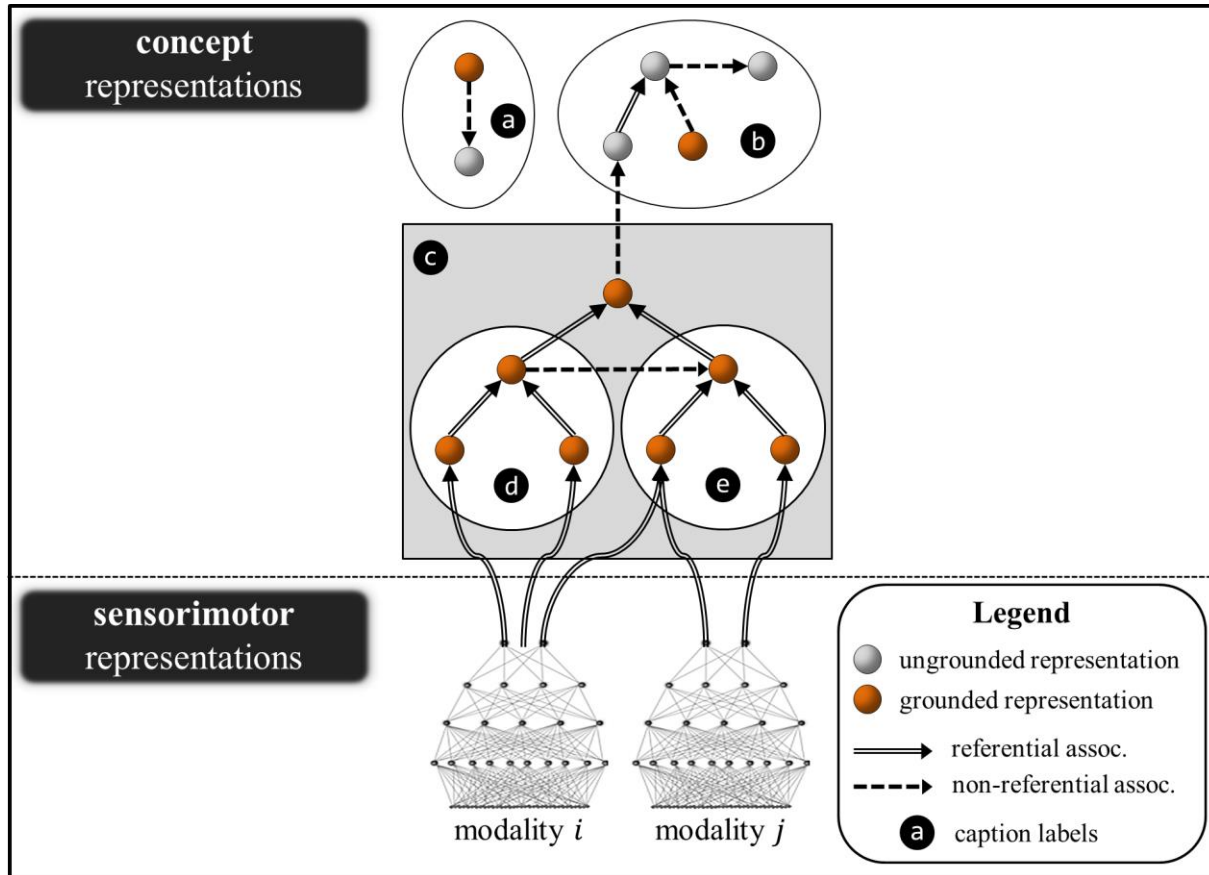


Figure 3. ES-Hybrid’s representational system. Representations are broadly divided into *concept* representations and *sensorimotor* representations. Several labeled examples are given in the diagram and described below: (a) contains an ungrounded concept representation (e.g., the unknown meaning of a word) with a *non-referential* association to a grounded representation (e.g., a word form); (b) contains an ungrounded concept representation with a *referential* association to another ungrounded representation and a *non-referential* association to a grounded representation (which may support analogical thinking); (c) contains a grounded concept representation corresponding to a *generalization* over several grounded concepts—(d) and (e) depict these concepts. Note that (d) has a non-referential connection to (e) and that (e) is multimodal. (Not all grounding, referential associations are depicted in the diagram.)

## Mental Representations

### *Sensorimotor Representations*

Sensorimotor representations are *non-symbolic* representations that emerge from the collective activity of a set of modality-specific, often hierarchically organized, *feature detectors*. These feature detectors are selectively receptive to “sense-able” aspects of an environment, for

example, olfactory or gustatory chemical signatures, acoustic vibrations, visible patterns of electromagnetic radiation, joint pressures, or nociceptive (i.e., pain) stimuli.

Each sensory modality may have drastically different characteristics and representational needs. For example, vision may require topographically organized photoreceptors, while olfaction may require combinations of highly specific chemoreceptors. Therefore, sensorimotor representations are typically segregated by sensory modality—that is, they will be primarily *unimodal*. However, multimodal representations can be formed from them through later associative bindings.

Sensorimotor representations are summaries of the *patterns of activation* that occur in sets of modality-specific feature detectors while sensing environmental stimuli. The resulting modal summaries characterize the most important low-level features of environmental stimuli. As such, sensorimotor representations are fundamental for many perceptual and simulation-related capabilities. They can also serve as the grounding constituents of grounded concept representations. Sensorimotor representations combine aspects of Harnad's (1990) iconic representations and Barsalou's (1999) perceptual symbols.

ES-Hybrid's sensorimotor representations possess the following properties:

- (1) They are *modal*, composed from the activity originating in sensory and/or motor systems.
- (2) They are *analogical*, bearing a resemblance to the things they signify.
- (3) They are *generative*, supporting modal mental simulations.
- (4) They are *perceptual*, capable of activating and being interpreted by perceptual systems.

### ***Grounded Concept Representations***

Sensorimotor representations are like the syllables or phonemes of modal meaning, while *grounded concept representations* are its words, phrases, and sentences. Like syllables and phonemes, sensorimotor representations are not directly meaningful, but they can become meaningful as the modal constituents of grounded concept representations. *Elementary* grounded concepts are formed when sensorimotor representations are bound together into (potentially) multimodal representations that can “stand-in” for specific sensory experiences. Binding occurs through the association of one or more modality-specific sensorimotor representations with a coordinating *amodal* representation that functions like a “convergence zone” (Damasio, 1989). These basic grounded concepts are viewpoint- and context-specific. They form the basis for resemblance-based discrimination (e.g., same/different judgments) and modal simulations.

Through the coordination of *generalization* and *predictive* processes, these “snapshots” of experience can lead to increasingly invariant and multi-part, grounded representations that correspond to particular objects, entities, situations, and events. And continued generalization can result in a hierarchy of categorical representations. Consequently, grounded concept representations exist in a *spectrum of generality* that extends from viewpoint-specific, sensory experiences to highly generalized categorical representations. Grounded concept representations can also grow into highly complex, hybrid representational structures—for example, structured, composite representations like Kosslyn et al.’s (2006) “object maps” (see Figure 4).

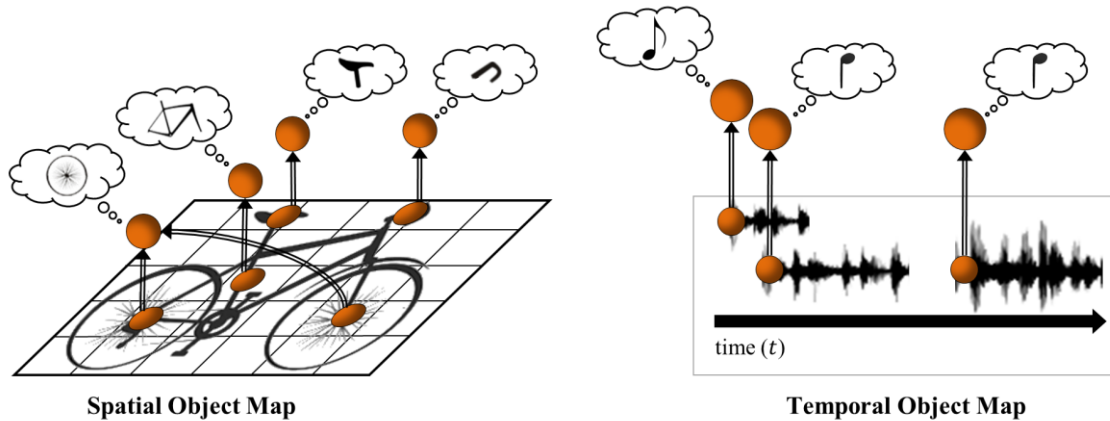


Figure 4. Grounded concept representations as object maps. A spatial object map (Left Panel) depicts the spatial extent of a bicycle with referential connections to grounded representations that correspond to its parts. A temporal object map (Right Panel) depicts a portion of a song with referential connections to grounded representations that correspond to individual notes.

All grounded concept representations are capable of being compared (e.g., by resemblance) and associated with other grounded concept representations (e.g., via referential and non-referential associations). They can also be mentally simulated due to their (direct or indirect) grounding in sensorimotor representations, which allows the partial re-enactment of these sensory experiences (e.g., how a particular sunset “looked”). Additionally, grounded representations functioning as concrete concepts support *categorical inferences* (i.e., identifying types from tokens) and the ability to instantiate (e.g., mentally simulate) specific instances of a category. Grounded concept representations reside within an associative network of referential and non-referential links (see Figure 3).

ES-Hybrid’s grounded concept representations possess the following properties:

- (1) They are *analogical*, bearing a resemblance to the things they signify.
- (2) They are *generative*, supporting modal mental simulations.
- (3) They are *grounded*, either directly or indirectly, in sensorimotor representations.



(4) They are *perceptual*,<sup>8</sup> capable of activating and being interpreted by perceptual systems.

These properties are discussed in more detail in the Appendix.

### ***Ungrounded Concept Representations***

*Ungrounded concept representations* are representational structures that are, at least *initially*, ungrounded. They are symbolic representations, which enables them to refer to any imaginable concept (concrete or abstract). However, taken as a whole, ungrounded concept representations are not “arbitrary” due to their associated indexical (contextualizing) grounded content.

Ungrounded concept representations result when predictive processes infer the existence of something that cannot, as yet, be identified with a grounded concept. For example, sensing a blur of motion out of the corner of one’s eye indicates that *something* exists that caused that visual blur, but it is not clear what that something may have been. Based on the hypothesized existence of this unidentified entity or event, predictive processes will

- (1) create an *amodal symbol*, and
- (2) create a *non-referential association* between the newly created amodal symbol and a grounded concept representation that corresponds to the sensory experience that initiated the causal speculation (e.g., the brief glimpse of an unidentified object in motion).

These amodal symbols serve as associative anchors that exert “pressure” on a cognitive system to establish their grounding. Their influence manifests in the selection of actions that seek to

---

<sup>8</sup> “Perceptual” grounded concepts implies that perceptual and conceptual systems can *share* (i.e., operate on) the same representations, and this terminology often appears in the grounded cognition literature (e.g., see Barsalou, 1999, 2016a; Goldstone & Barsalou, 1998; Haimovici, 2018). Therefore, I use the term “shared” when discussing this representational property in the Appendix.

ascertain the identity of these unknowns. The resulting actions may focus an agent's attention or re-orient its body towards sensory stimuli of interest (e.g., towards a perceived motion). In other words, the awareness of these unknowns can compel agents towards active exploration.

The identity of these unknowns can often be determined either empirically (e.g., via exploratory external actions) or through speculative reasoning (e.g., via exploratory internal actions). In either case, if a real or hypothesized causal entity or event is determined, then offline cognitive processes will create a *referential association* to that causal entity or event. Crucially, without these initially ungrounded amodal symbols, we would have nothing to organize our empirical evidence and speculative reasoning around. In other words, they serve as “scaffolding” that support the eventual acquisition of modal meaning.

In most cases, ungrounded concept representations are *eventually grounded*. This occurs when referential associations are established (learned) with grounded concept representations. However, if an amodal symbol refers to an abstract concept, it will remain permanently detached from an agent's sensory systems. In some cases, concrete metaphors may exist (e.g., “time flies like an arrow”), allowing *non-referential associations* to be created between these *ungroundable* concept representations and grounded concept representations.

Though permanently ungrounded, the amodal symbols for abstract concepts can, by non-referential associations with other modal and amodal representations, carry contextual and relational information (e.g., supporting distributional semantics; see Chapter 2). They can also establish referential associations with other amodal symbols (e.g., **X IS A Y**). However, they will always lack “modal meaning” since they are incapable of being grounded, and, by extension, they cannot support non-metaphorical, modal simulations.

ES-Hybrid's ungrounded concept representations possess the following properties:

- (1) They are *symbolic*, bearing no resemblance to the things they signify.
- (2) They are *ungrounded*, having no direct or indirect referential associations with sensorimotor representations.
- (3) They are *non-generative*, unable to support modal mental simulations.
- (4) They are *non-perceptual*, incapable of activating and being directly interpreted by perceptual systems.

Notice that most of these properties are negations: They are not grounded; They are not generative; They are not perceptual. Consequently, their referents are also largely *unconstrained*. This makes ungrounded concept representations incredibly powerful: they can represent almost anything.

### ***Referential and Non-Referential Associations***

Associations (i.e., “links”) between representations are broadly characterized as either *referential* or *non-referential*. Note that while it is often convenient to discuss these associations in terms of graph-theoretic concepts (i.e., nodes and links), this should not be interpreted as a constraint on implementations. Any mechanisms by which two representations can be associated—for example, synaptic connections or cross-frequency (e.g., theta-gamma) coupling—is fair game.

Referential associations are thus named because they directly support the establishment of a representation's referent; that is, they establish what it corresponds to or signifies in the world. For example, “this smell *corresponds to* garlic” or “this feeling *corresponds to* pain” or “this mathematical symbol *signifies* the ratio of a circle's circumference to its diameter.” For

elementary grounded representations, this relationship takes the form of a constitutive association with sensorimotor representations; that is, sensorimotor representations directly comprise those representations. For categorical concepts, referential associations can take the form of “is a” (i.e., set membership) relationships; for example, “Mr. Muggles *is a* Pomeranian” or “a Pomeranian *is a* type of dog” or “ $\pi$  *is a* transcendental number.” Here, a categorical instance (i.e., “token”) signifies a category (i.e., “type”). In both cases, the representational nature of the referential associations is identical. However, their semantic interpretations—which depend on the cognitive processes that use them—may subtly differ.

Non-referential associations broadly encompass all non-grounding associative relationships. These could include causality (e.g., “gravity *caused* the apple to fall”); indexical correlations (e.g., “smoke indicates the presence of fire”); sequential orderings (e.g., “**A** occurs before **B** in the English alphabet”); relational operators (e.g., “ $10 < 11$ ”), etc. In general, any association that does not establish an iconic or symbolic signifier/signified relationship between two representations can function as a non-referential association, and the nature of an agent’s offline cognitive processes will largely determine which non-referential associations are needed and how they are established.

Non-referential associations between abstract and concrete concepts allow nebulous (abstract) ideas to be conceptualized through *analogy* with more tangible concrete concepts. This contrasts with referential associations between concept representations which can allow ungrounded representations to *become* concrete concepts through “indirect grounding.”

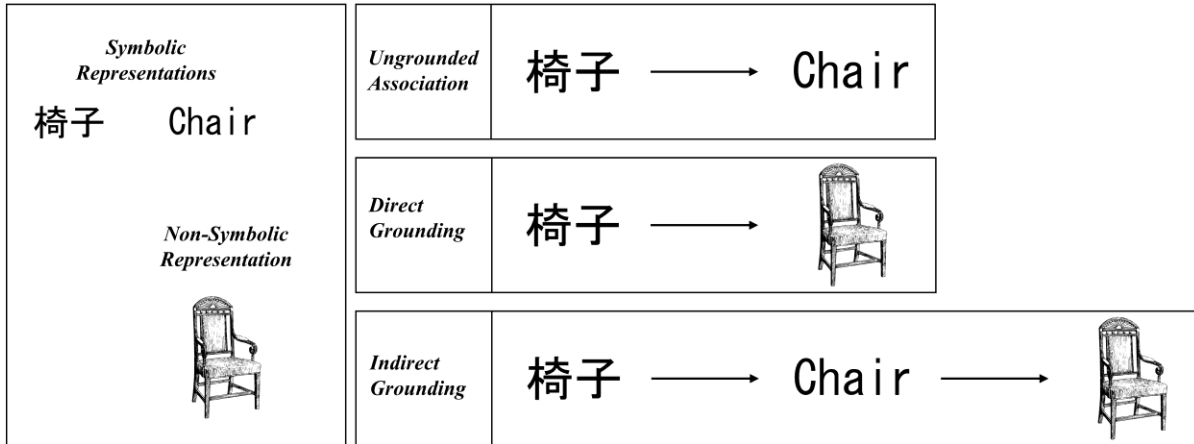


Figure 5. Intuitive depiction of direct and indirect grounding. Arrows indicate referential associations. Referential associations between purely symbolic representations are not sufficient to establish grounding. Only referential associations that are terminated in a non-symbolic representation can ground a symbolic representation. Symbolic representations can be indirectly grounded if a chain of referential associations connects the symbolic representation to a non-symbolic representation.

### ***Grounding***

A concept representation is *minimally grounded*<sup>9</sup> if and only if there exists a chain of referential associations from it to a sensorimotor representation. A concept representation is *directly grounded* if and only if there exists a single-link referential chain (i.e., a direct association) from it to one or more sensorimotor representations. And a concept representation is *indirectly grounded* if and only if it is not directly grounded, and a referential chain exists (of length two or more) between it and one or more sensorimotor representations. Figure 5 attempts to illustrate these grounding relationships using the more intuitive notions of symbolic and non-symbolic representations. Figure 3 depicts these relationships within the context of ES-Hybrid’s conceptual system.

---

<sup>9</sup> “Minimally grounded” was chosen instead of “grounded” to reinforce the notion that grounding is not marked by a singular “grounding” event, but rather the gradual establishment of referential associations between conceptual representations and sensorimotor representations.

Grounding is often portrayed as a Boolean (true or false) property of representations—like establishing a grounding connection in an electrical circuit. In this view, a representation is either grounded, or it is not. However, such metaphors are misleading. Instead, I contend that grounding is a continual activity that involves incrementally establishing referential connections between concept representations and their corresponding sensorimotor representations. The more referential connections (i.e., referential chains) to sensorimotor representations, the more grounded a concept representation can be said to be. For example, a concept representation that is grounded in visual, auditory, and tactile sensorimotor representations can be said to be *more grounded* than one that is only grounded in visual sensorimotor representations (*ceteris paribus*).

## **Cognitive Processes**

### ***Generalization***

Automatic, bottom-up generalization processes<sup>10</sup> use the similarities between sensorimotor representations and grounded concept representations to *cluster* related representations. This can support the formation of more general concept representations by distilling the commonalities over those exemplars. For example, generalization processes might organize the *viewpoint-specific* sensory experiences of a single cat into a *viewpoint-invariant* representation for that cat. Similarly, grounded representations corresponding to many individual cats can be generalized into concrete concepts for cats at large.

---

<sup>10</sup> The generalization processes I refer to in this section are unconscious, bottom-up, and resemblance-based. This contrasts with conscious and deliberative generalization processes that can construct more general concepts based on conscious feature inspection and categorization. The taxonomic classification of organisms into kingdoms, phyla, families, genera, and species by phylogenetic characteristics is an example of the latter (which will not be covered).

Gradually, the representations generated by these processes can become selectively sensitive to the activity of some grounded concept representations but not others (cf. Harnad's categorical representations). In this way, grounded concepts for *subordinate* (e.g., apple), *basic* (e.g., fruit), and *superordinate* (e.g., food) categories can be formed (see *Cognition and Categorization*, 1978). The specifics of these generalization processes and the conceptual depth of the learned ontology will depend on the needs of the agent and the tasks it endeavors to complete. Generalization processes can work in conjunction with perceptual and predictive processes to identify part-whole relationships in order to form structured representations, such as "object maps" (Kosslyn et al., 2006).

### ***Prediction***

ES-Hybrid's predictive processes are imagistic and associative processes that make inferences about the objects, entities, situations, and events that *active representations* signify. For example, they may be used to infer causal relationships (e.g., a spilled drink *caused* the puddle on the kitchen table) or predict action consequences (e.g., neglecting to water a houseplant *may cause* it to wilt and die). These predictive processes support the creation of referential and non-referential associations, and they can be used to identify unknown referents. More generally, the primary function of mental simulation may be the generation of experience-based predictions (Moulton & Kosslyn, 2009). And numerous researchers have noted a relationship between prediction, perception, and mental simulations (Barsalou, 2009; Clark, 2013; Jeannerod, 2001).

### ***Mental Simulation***

Mental simulation proceeds by iteratively re-activating grounded representations in a top-down fashion. This process terminates when modal sensorimotor representations have been reactivated

and used by generative processes to create modal simulations. Generative processes create modal simulations by inducing patterns of activation in an agent's sensory and motor systems that correspond to grounded concepts. These simulations are then integrated into "virtual scenes" within an agent's internal environment.

Mental simulation is a largely learned capability that is directly supported by the acquisition of sensorimotor representations and referential (grounding) associations. Loosely speaking, the more referential associations exist between a concept representation and its corresponding sensorimotor representations, the greater one's capacity to mentally simulate those concepts. Furthermore, since grounded representations are required for mental simulation, the extent of one's learning about a (concrete) concept might be quantifiable through one's capacity to mentally simulate it (Barsalou, 1999, sec. 2.4.3). Conversely, the failure to adequately simulate an object, event, or the consequences of one's actions indicates a predictive failure that can be exploited by a cognitive system—for example, to orient one's attention towards "surprising" environmental stimuli<sup>11</sup>.

### ***Action Selection***

Offline cognitive activities culminate in the activation and/or generation of the representational precursors of action selection. These representations may correspond to an agent's current beliefs, desires, and intentions (Bratman, 1987), elements of its ongoing plans, or aspects of its current situation (among other things).

---

<sup>11</sup> This idea is exploited in Chapter 5 to direct an agent's attention during perceptual learning. The same mechanism could be used to support procedural learning as well.



ES-Hybrid is generally agnostic to the details of this action selection mechanism; however, it is crucial that selected actions should serve as high-level, goal-oriented directives (e.g., hitting a tennis ball with a forehand swing) rather than low-level, actuator commands (e.g., the excitation of collections of muscle fibers). These actions should “predispose” a cognitive system towards a particular mode of *situated online control* (see the next section), rather than dictate the execution of individual motor commands. That is, selected actions should specify *what* should be done rather than *how* it should be done.

Actions can be selected for internal (covert) or external (overt) execution. Internal actions affect an agent’s internal environment, while external actions affect its external environment. Internal environments are composed of affectable and introspectable mental states. External environments are composed of sense-able objects and events that are assumed to exist beyond an agent’s private mental states.

### ***Action Execution***

Once an action is selected, its *overt* execution is fulfilled through a process of *situated online control*. This control mechanism proceeds in parallel and largely independent of offline cognitive processes, and it continues until another action is selected. This situated control subsystem is envisioned as being largely non-representational and reactive (cf. the subsumption architecture; Brooks, 1986, 1990). It is tightly coupled to the environment, continually influenced by incoming sensory stimuli and the environmental effects of ongoing action execution.

In dynamic, rapidly changing environments, such a parallel and largely independent mechanism of situated online control is essential. Offline cognitive processes are often too slow to react to the changing demands of such environments. Therefore, ES-Hybrid incorporates

multiple control loops. The offline control loop's *raison d'être* is to periodically (re-)orient the online control loop's situated activity towards some proximal intention.

One final note: it has been hypothesized that the *outflowing* motor signals sent to actuators during action execution produce “corollary discharges” (Crapse & Sommer, 2008; Jeannerod et al., 1979; Sperry, 1950). Corollary discharges can result in *inflowing* copies of outflowing motor signals that can be used to influence an agent's sensory and perceptual systems.<sup>12</sup> In particular, they can be used to predict the anticipated sensory consequences of motor command execution (e.g., using a forward model; see D. M. Wolpert et al., 1995) and adjust the patterns of activation in sensory systems accordingly. Such predictive sensory feedback is a form of low-level motor simulation; therefore, it has been included in ES-Hybrid's conceptual framework for completeness.

### **Amodal Representations for Grounded Cognition**

Amodal representations have been mischaracterized by focusing, almost exclusively, on a single class of amodal representations, namely those that appear in formal symbolic systems like mathematics, logic, and human languages. For example, Barsalou stated

[a]modal symbols bear an important relation to words and language. Theorists typically use linguistic forms to represent amodal symbols... [and] symbolic thought is assumed to be analogous in many important ways to language. Just as language processing involves the sequential processing of words in a sentence, so [amodal] conceptual processing is

---

<sup>12</sup> Depending on the context, corollary discharges are referred to as “efference copies” (Von Holst, 1954), though these concepts are subtly different.

assumed to involve the sequential processing of amodal symbols in list-like or sentence-like structures. (Barsalou, 1999, p. 579)

This characterization assumes that amodal representations require rule-based, symbolic manipulations for them to serve a purpose within a cognitive system. However, I contend that this assumption is fallacious, and it is largely based on their historical uses rather than their potential.

While humans may acquire amodal representations in support of activities such as mathematics and language, I contend that such human-centric use cases are derivative of a more fundamental class of amodal representations that are shared with non-human animals. Specifically, I contend that amodal representations are generated by an agent's top-down, predictive processes when the existence of an *unknown* entity, object, concept, or event is inferred. These amodal representations can facilitate numerous cognitive functions without resorting to classical rule-based symbolic manipulation. Moreover, these amodal representations can directly support grounded cognition, rather than standing in opposition to it. In the subsections that follow, I review some of the use cases for amodal representations that do not require rule-based symbolic manipulation.

### ***Unknown Referents***

Consider the following scenario: Alice, our agent, is taking a stroll through the woods, when she hears the sound of leaves rustling behind her. Startled by the unexpected sound, she turns in the direction of its source in search of a cause. One plausible cognitive account of these events is the following:

- (1) Alice heard a sound behind her, which she recognized as the sound of leaves rustling.
- (2) Alice predicted that this rustling of leaves had a cause, though this cause is initially unknown (but presumably knowable).
- (3) As a result of this prediction, Alice re-oriented her body to acquire additional evidence (i.e., sensory stimuli) in an attempt to identify the cause of the rustling leaves.

Figure 6 depicts a knowledge structure (in two different formats) that could support these activities by representing the existence of a hypothesized, but initially unknown, cause of the rustling leaves.

Figure 6, Panel (a) shows a purely symbolic (amodal) representation of this knowledge structure. The variable  $c_1$  is introduced to represent the unknown cause of the leaves rustling. A **Caused** relation associates  $c_1$  to what it caused—i.e., the leaves to rustle, which is depicted using the atomic symbol **LEAVES\_RUSTLING**. The **Caused** relation, when combined with these specific arguments, is intended to convey the idea, “the rustling of the leaves was *caused by* some currently unknown (but likely knowable) entity or force.”

(a)

Caused(Cause =  $c_1$ , Effect = LEAVES\_RUSTLING)

(b)

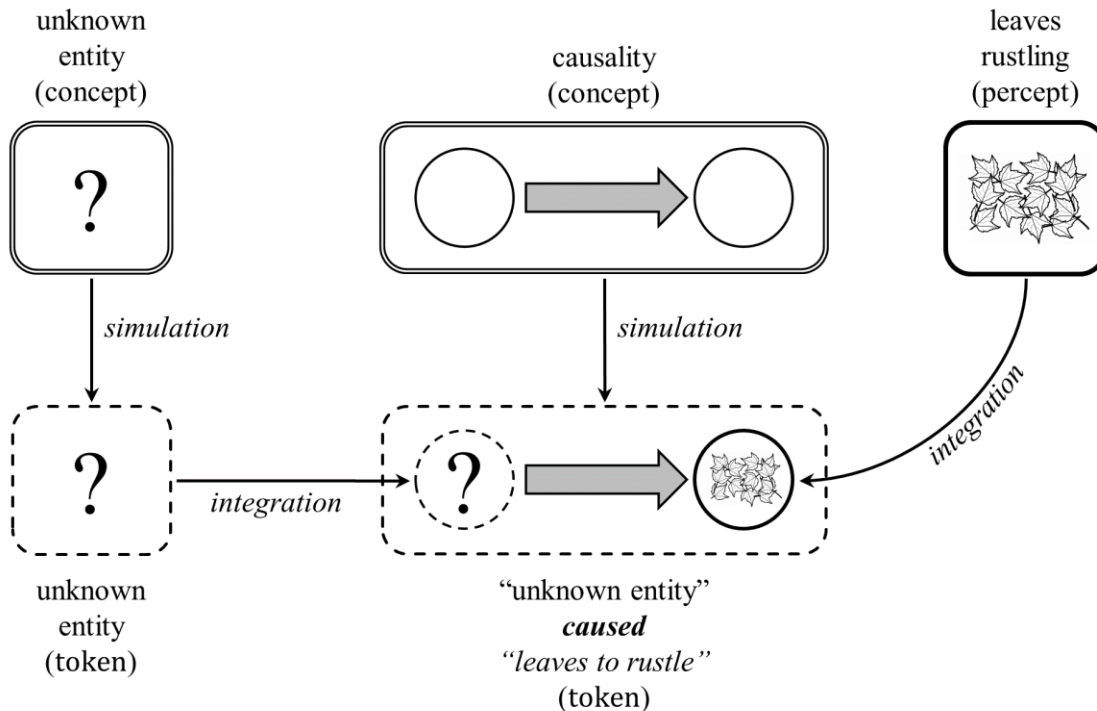


Figure 6. Depiction of a hypothesized unknown entity. Panel (a) depicts a purely symbolic representation corresponding to an unknown entity that is hypothesized to have caused the event of leaves rustling. Panel (b) depicts the same knowledge using a simulation-based knowledge structure. (Rectangular regions with double-line borders represent simulators, rectangular regions with dashed borders represent simulations, and rectangular regions with thick borders represent modal percepts corresponding to incoming sensory stimuli.)

Figure 6, Panel (b) shows the same knowledge encoded using an integrated set of simulations and modal representations<sup>13</sup>. A simulator (see Chapter 2, “Embodied, Simulation-Based Cognition and Perceptual Symbol Systems”) corresponding to Alice’s concept of “unknown entity” generates a simulation capable of designating the existence of an unknown

<sup>13</sup> The depictive style used in Figure 6, Panel (b) was based on Barsalou’s (1999) figures depicting the integration of perceptual symbols into composite representations (e.g., see Barsalou, 1999, fig. 5).

entity (or force). Another simulator, corresponding to Alice’s concept of “causality,” generates the simulation of a causal event representing “some unknown entity caused the leaves to rustle.” Note that this causal event simulation is itself composed of the previously constructed “unknown entity” simulation (i.e., the cause) and a modal percept corresponds to the “leaves rustling” (i.e., the effect). Its function is to associate the “unknown entity” with the event of the leaves rustling via a causality relationship.

Question: Is the simulation generated for the “unknown entity” concept, depicted in Figure 6, Panel (b), a modal or an amodal representation? Let us consider the necessary properties of modal representations (discussed earlier in this chapter).

- (1) Is it composed of *modality-specific content* (i.e., content originating in the agent’s sensory or motor system)?
- (2) Does it bear an *analogical* relationship with its referent (i.e., does it resemble, share observable properties, or a structural correspondence with its referent)?
- (3) Is it *grounded* (i.e., are there informational conduits in place that could establish a correspondence between it and some observable, environmental stimuli)?
- (4) Is it *sharable* between the agent’s perceptual and conceptual systems?

Answering “yes” to any of these questions is exceedingly difficult.

The concept of an “unknown entity” is characterized by its lack of a specific identity and the absence of tangible properties. These characteristics make it challenging to determine *any* modal (sensory or motor) content that could be constitutive of this concept or one of its

simulations. Furthermore, the lack of known properties undermines its analogical relationship with its referent, which *does*, in fact, have properties. Alice simply does not know what they are. Given that Alice has yet to observe the entity that caused the leaves to rustle, she has yet to learn any correspondence between the “unknown entity” representation and the thing in the world to which it refers. In other words, the representation appears to be *ungrounded*. Note that the “rustling of the leaves” percept that resulted in the generation of the “unknown entity” representation does not refer to (i.e., it is not “about”) the unknown entity that may have caused the leaves to rustle. Rather, the rustling leaves are *indexical* of the unknown entity’s existence in the same way that smoke indicates the likely existence of a fire, or someone’s finger points to something behind you. (I will return to this relationship between indexical context and “unknown referents” shortly.) Finally, the “unknown entity” representation does not seem to contain any content that would be usable by a perceptual system, in that it seems devoid of perceivable content, components, and properties. In summary, this simulation appears to be amodal.

One possible objection to the representation in Figure 6, Panel (b) is that the “unknown entity” simulation is better modeled as an “unbound” portion of the “causality” simulation. The idea being that the *absence* of a specific causal element in the simulation could be interpreted by Alice’s conceptual system as the *existence* of something to which Alice knows nothing about. There are several problems with this interpretation. First, the meaning of a missing element in a coordinating simulation is ambiguous. It could mean “nothing,” or it could mean “anything.” Second, the agent’s perceptual system should be able to simulate an “unknown entity” (or any number of such entities) in the absence of a mediating composite simulation; however, without a coordinating simulation or “modal frame” (Barsalou, 1999, sec. 2.5) there is no way to depict

this absence. Third, Alice’s representation for the “unknown entity” must be able to serve as an associative anchor (i.e., a container, scaffold, or binding agent) that supports the gradual acquisition of information about that entity. A coordinating simulation can provide the necessary context, but a void in a coordinating simulation is insufficient to serve as the necessary associative anchor. Finally, the same “unknown entity” should be referenceable from any number of coordinating simulations and contexts, even in the absence of known properties or contextual clues. If the only mechanism available to the system is the absence of content, then there is no way to indicate that “this absence” refers to the same thing as “that absence.”

For all of the reasons given above, I contend that coordinating *amodal* representations are needed to represent *unknown referents*. These amodal symbols serve as cognitive placeholders that “point to” (i.e., they are indexical of) the existence of something of interest without initially grounding it or specifying its meaning. This leads to a critical observation: these amodal representations generally exist in conjunction with an *indexical* representation. That is, they are composite, hybrid (modal/amodal) representations. In effect, these two representations are pair-bonded. The indexical representation enables an agent to orient its body and cognitive processes in order to gather additional information about the unknown referent; it serves as a contextualized clue indicating the conditions under which the unknown entity may be observable (e.g., when the leaves are rustling). The amodal representation scaffolds the gathering of that information and serves as an opaque reference to the unknown entity wherever referential consistency is needed. Neither of these representations could stand alone without undermining the intent of the representational structure.



### *Cognitive Indirection*

Consider the following example. Suppose that our agent, Alice, is confronted by a man, named Bob, that asks whether she has heard about the **WUMPUS-ALPHA-PRIME**. Uncertain, Alice speculates about the possible nature of the **WUMPUS-ALPHA-PRIME** (based on linguistic similarity). Perhaps **WUMPUS-ALPHA-PRIME** refers to a star system (similar to *Alpha Centauri*), or the first (*prime*) planetary body in that star system; or the first derivative (in *prime* notation) of a variable called  $WUMPUS_{\alpha}$ ; or a character in a video game designed by Gregory Yob<sup>14</sup>?

Deciding that these explanations are unlikely, Alice interrogates Bob for more details. Bob tells Alice that he last saw the **WUMPUS-ALPHA-PRIME** on a recent expedition to the arctic circle. He goes on to say that the **WUMPUS-ALPHA-PRIME** is a massive creature, nearly twice the height of an average person when standing on its hind legs. Its skin is covered by thick, white fur; it has huge, sharp claws; and a distinctly *bear-like* body. After a series of guesses, and some creative mental imagery, Alice believes she has converged on the identity for the **WUMPUS-ALPHA-PRIME**: it must be a *polar bear*. However, before Alice can inform Bob of his terminological mistake, he continues. Bob tells Alice that the **WUMPUS-ALPHA-PRIME** has nine tentacles growing out of its back, which it uses to catch its prey. And that it speaks fluent Hungarian, but with a slightly British accent.

---

<sup>14</sup> Gregory Yob is the software developer who created the text-based adventure game *Hunt The Wumpus*, which subsequently inspired the classical grid world agent environment “Wumpus World” mentioned in Russell and Norvig’s book *Artificial Intelligent: A Modern Approach* (Russell & Norvig, 1995/2010)

The mental gymnastics that occurred in Alice's mind while trying to ascertain the identity of the **WUMPUS-ALPHA-PRIME** is an example of what I am referring to as *cognitive indirection*<sup>15</sup>. It is the ability to rapidly adjust a mental representation's referential and non-referential associations, while maintaining any dependent associations (i.e., the other representations that refer to it). In this way, a representation's properties, grounding, and intentionality can be fluidly changed while simultaneously maintaining other previously established associative relationships. I will elaborate on this further in the context of the previous example.

I contend that upon learning of the existence of the **WUMPUS-ALPHA-PRIME**, an initially ungrounded, amodal representation was created (or otherwise allocated) in Alice's mind to represent this unknown referent. The need for this representation was inferred (predicted) from the initial situational and linguistic context. This amodal representation *internally symbolizes* the **WUMPUS-ALPHA-PRIME** (an initially unspecified and ungrounded concept), allowing it to be associated with a specific situational context, as well as a modal representation for an auditory word-form ('wʌm pəs 'æɪ fə praɪm) that *externally symbolizes* whatever object, entity, or event the **WUMPUS-ALPHA-PRIME** refers. This auditory word-form is a separate concept in its own right: a linguistic token grounded in a sequence of modality-specific sensorimotor representations (auditory stimuli). Critically, this word-form is *indexical* of the **WUMPUS-ALPHA-PRIME** concept. Whenever Alice hears this word-form in the future, the **WUMPUS-ALPHA-PRIME** concept in Alice's mind will likely be activated.

---

<sup>15</sup> *Indirection* refers to a concept in computer programming where a variable is used as a reference to another variable that contains a value, rather than referring to that value directly.

Following the creation of this hybrid (amodal/modal) composite representation, a brief period of speculation occurred during which Alice *internally* explored possible referents for the **WUMPUS-ALPHA-PRIME**. For each of these, a referential association was established between the mental representation for that hypothesized referent and the amodal representation for the **WUMPUS-ALPHA-PRIME**. These transient associations were then discarded (i.e., the associative bounds between these mental representations were dissolved). Despite these structural changes, the association between the word-form for the **WUMPUS-ALPHA-PRIME** and the amodal representation corresponding to that concept remain unscathed. I contend that Alice’s ability to engage in such speculative reasoning requires a form of cognitive indirection that is facilitated by a mediating amodal representation. This situation is depicted in Figure 7.

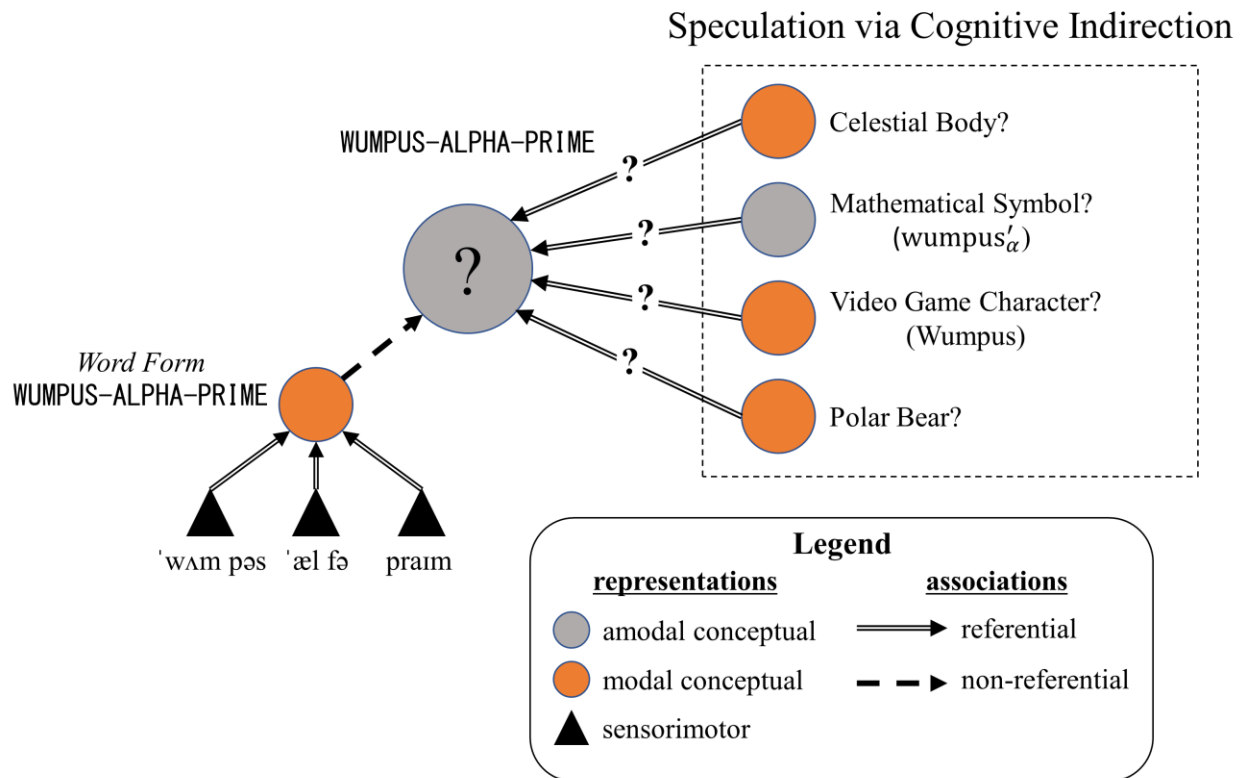


Figure 7. Illustration of cognitive indirection.

Finally, Alice abandoned this *internal* exploration in favor of *external* exploration. She interrogated Bob for additional information, resulting in a transitory belief that she had identified the **WUMPUS-ALPHA-PRIME** as a polar bear. Then, she immediately discarded this unfounded belief when contradictory information emerged (e.g., tentacles and such). Notice that even though Alice dismissed her belief that the **WUMPUS-ALPHA-PRIME** was identical to a polar bear, the new property attributions she learned about it (e.g., its thick, white fur; huge, sharp claws; and bear-like body) must remain. In fact, it is entirely reasonable to suspect that Alice might use the concept of polar bears as the basis for analogical reasoning about the **WUMPUS-ALPHA-PRIME** since it is unlikely that she (or anyone else) will directly observe the **WUMPUS-ALPHA-PRIME** in its natural habitat.

### ***Multimodal Representations***

Amodal representations provide a simple and flexible way of binding multiple, modality-specific representations into composite, *multimodal* representations. For example, elementary grounded conceptual representations (described earlier in this chapter) can be formed by binding together multiple sensorimotor representations via referential associations. Once bound, these composite representations can “stand-in” for specific sensory experiences.

This mechanism is largely consistent with the “hub-and-spoke” model of mental representations (see Patterson et al., 2007; Ralph et al., 2010, 2017). This model states that while concepts are largely represented in modality-specific cortical regions of the brain, the inter-modal interactions between those regions must be at least partially mediated by a trans-modal “hubs.”

These amodal hubs could also be conceptualized as instances of Damasio's (1989) *convergence zones*, which are coordinating neural assemblies that support the routing of information, the binding of neural patterns of activation, and the cross-modal re-activation of sensorimotor representations. Damasio described convergence zones as amodal because they “do not map sensory or motor activity in a way that preserves feature-based, topographic and topological relations of the external environment as they appear in psychological experience..., [and] they are uninformed as to the content of the representations they assist in attempting to reconstruct.” (1989, p. 46). Convergence zones have been characterized as “pointers” (e.g., see Lalleo & Dominey, 2013; Tyler et al., 2004) since they indicate where information is stored, rather than representing it directly.

### ***Designation and Disambiguation***

Frequently, resemblance—e.g., what an object looks like or sounds like—is insufficient to establish an object's (or event's, situation's, etc.) unique identity. For example, in Figure 4 (left panel) the bicycle's front and rear wheels appear to be identical (in isolation), and they are associated with the same conceptual representation. Nevertheless, they have distinct identities despite having identical appearances.

*Designating* (see Barsalou, 1999, sec. 2.2.3) that a mental representation refers to a specific instance among a class of identical objects requires additional information, and that information can include highly fluid—and relatively arbitrary—contextual and historical details. For example, returning to the bicycle in Figure 4 (left panel), the *front* wheel and *rear* wheel can be easily identified as such while they are still attached to its frame. However, if both wheels are removed from the bicycle, ancillary details are needed to establish their identities. They continue

to be *this* wheel and *that* wheel (i.e., distinct wheels), and if someone were to ask—“Was *this* wheel the front wheel or the back wheel?”—it may be possible to answer that question.

However, one must remember details such as “the front wheel was placed on the work bench” or “the rear wheel was placed on the floor.”

Tracking an object’s persistent identity—despite ever changing contextual and resemblance details—requires that something represent for that specific object, and that the appropriate ancillary, disambiguating details are associated with it (e.g., last known location). Moreover, even if those associated details are lost (e.g., the object’s current location), we can still recognize that a distinct object with that identity exists (or did exist)—regardless of whether we can ascertain its identity.

Amodal representations excel at representing distinct objects, entities, situations, and events. A unique amodal symbol can be allocated to a specific instance of a class of objects, and, even if *all* details associated with that instance change (e.g., resemblance, context, history), it can still function as *that* object. Moreover, the details associated with two objects could be identical (e.g., same appearance, context, history, etc.), any they can still be identified as being distinct. As a corollary to this property, amodal symbols can serve as markers (i.e., designators) for “whatever thing happens to be here” within a scene, cognitive map (Tolman, 1948), or object map (Kosslyn, 1994; Kosslyn et al., 2006).

### ***Active Perception***

Initially ungrounded, amodal symbols are not fringe phenomena. They are the pervasive byproducts of top-down, predictive, cognitive processing. These ungrounded representations exert “pressure” on a cognitive system that encourages the selection of actions that seek to

ground them. These actions focus our attention. They orient our bodies towards the unknown. And they compel us towards active exploration and speculation in order to ascertain the sources of these unknowns. In most cases, these initially ungrounded amodal symbols can be *eventually grounded* through learned associations with the tangible.

### ***Eventual Grounding***

Implicit in most definitions of grounding is the idea that grounding is established through bottom-up generalization over non-symbolic representations (cf. Barsalou, 1999; Harnad, 1990). However, I contend that grounding can also be initiated by *top-down* processes. In these cases, an *initially ungrounded* amodal representation—corresponding to an unknown, hypothesized referent—is generated by a predictive process. This ungrounded representation can be later grounded by establishing a referential association between it and a grounded concept representation.

A common example of this occurs during language acquisition (or other modes of formal education) when a learner is presented with an unfamiliar word. The student knows that the word must refer to something, but its referent is initially unknown (and the set of possible referents may be largely unbounded). Through continued instruction and inquiry, the word's meaning will hopefully be ascertained by the student. In the case of concrete concepts, this initially ungrounded amodal symbol may eventually be grounded by creating a referential link to a corresponding grounded concept representation. However, if the word signifies an abstract concept, grounding may not be possible.

## **On Language**

Language comprehension and production are complex cognitive functions that I cannot do justice to here. However, I contend that elements of the theory advanced here could be elaborated on to support a theory of language acquisition. For example, upon hearing or seeing an unknown word, a predictive process can create an amodal symbol for that inferred, underlying concept. This initially ungrounded representation can then be associated (using a non-referential association) with the (visual or auditory) word form for that concept. The word form serves an indexical function (in the Peircean sense) for the underlying concept, indicating the situational contexts in which that concept may be pertinent. After repeated exposure to that word form and its associated situational elements, an individual could distill the invariant environment components that characterize that concept.

ES-Hybrid is consistent with the notion that language comprehension could employ a combination of perceptual/conceptual knowledge, embodied simulations, and language statistics (Barsalou et al., 2008; Louwrese, 2011, 2018; Louwrese & Jeuniaux, 2008). In particular, non-referential associations between word forms could support context-based sources of meaning (e.g., distributional semantics). That said, the decision to avoid language-specific processes is not only a practical consideration, but it reflects my belief that a full account of non-linguistic processes is a prerequisite to understanding the relatively human-centric experiences entailed by language use. As Barsalou et al. (2008) stated, it is likely that conceptual systems evolved primarily to process non-linguistic stimuli (e.g., perceptual, motor, and introspective experiences), and these non-linguistic experiences are likely more fundamental to human cognition than the processing of words. In general, I agree with Newell when he wrote,



“Language should be approached with caution and circumspection.... I will take it as something to be approached later rather than sooner” (Newell, 1994, p. 16).

## Chapter 4

### Learning Intelligent Decision Agent (LIDA)

There is nothing so practical as a good theory. (Lewin, 1951, p. 169)

Much of the current research in embodied, simulation-based cognition (ES) is limited to high-level theories. Classical cognitive theories, on the other hand, possess a high-level theory as well as more detailed cognitive models and computational mechanisms. In order to compete with classical cognitive theories, ES must implement complete agential systems<sup>1</sup> that make manifest the consequences and predictions of those theories in software.

ES-Hybrid—the hybrid ES theory developed in this manuscript—is too underspecified and abstract to be used directly for implementing complete agential systems. It needs to be conceptualized within a cognitive architecture to fill in its missing details. Towards this end, I will integrate many of the fundamental ideas from ES-Hybrid into the Learning Intelligent Decision Agent (LIDA; Franklin et al., 2016) cognitive architecture.

I will begin this enterprise with a brief introduction to the LIDA cognitive architecture. Subsequent chapters will describe how specific ES-Hybrid functionality can be conceptualized within LIDA. Chapter 5 will focus on implementing grounded representations, modal simulations, and multimodal perception. Chapter 6 will focus on *action-based* mental simulation

---

<sup>1</sup> Many thought leaders, including Newell (1973), Brooks (1991b), and S. Wilson (1991), have argued that it is essential to model complete agential systems rather than isolated competences. While the complexity of understanding and building intelligent systems might compel us to focus on more manageable sub-problems (e.g., vision or natural language processing), the danger is that the resulting specialized models may be irrelevant in the context of a fully functional, autonomous agential system. By modeling complete systems (from sensing to acting), we reduce the risk that a system's individual competences are incompatible with each other and the system at large.

and motor cognition. And Chapter 7 will describe the fundamental operations of mental imagery and simulation-based, epistemic (knowledge-generating) processes.

## Overview

LIDA (Learning Intelligent Decision Agent; see Franklin et al., 2016) is a biologically inspired cognitive architecture<sup>2</sup> that strives to be a “unified theory of cognition” (Newell, 1994) that is capable of modeling many, if not all, cognitive activities and processes. Cognition, in this sense, broadly encompasses every mechanism of mind, including (but not limited to) perception, motivations, action selection, motor control, attention, learning, metacognition, language, sense of body and self, and mental simulation. In addition to supporting the analytical modeling of cognitive processes, cognitive architectures also facilitate the creation of complete agential software systems; therefore, cognitive architectures are useful for both analytical and synthetic approaches to understanding minds (see Franklin, 1995, pp. 9–10).

While LIDA is a *biologically inspired cognitive architecture* (BICA), it does not attempt to model brains<sup>3</sup>—it models *minds* (see Chapter 2). Minds, in this context, are defined as “control structures for autonomous agents” (Franklin, 1995, p. 412). Where an *autonomous agent* is any natural or engineered system that is “situated within and a part of an environment that senses that environment and acts on it, over time, in pursuit of its own agenda and so as to effect what it senses in the future” (Franklin & Graesser, 1997, p. 25). According to this definition,

---

<sup>2</sup> See Kotseruba and Tsotsos (2018) for an excellent survey of many of the cognitive architectures that have been developed from the early 1980s to present.

<sup>3</sup> LIDA’s modular organization is not directly based on the neural architecture, or the gross anatomy, of brains, but on the functional relationships that appear to exist between various cognitive processes (e.g., inferred from psychological experiments and functional analysis) and the mental representations they support. LIDA does not make any assertions about the modular organization of brains.

autonomous software agents differentiate themselves from standard “programs” by their situated and embedded relationship with an environment and their selection of actions that further their own agenda.

LIDA differentiates itself from other cognitive architectures in several ways that make it ideal for implementing a hybrid account of ES:

1. LIDA is a hybrid cognitive architecture (see Kotseruba & Tsotsos, 2018, sec. 3) that features symbolic and non-symbolic representations, as well as non-representational processes.
2. LIDA has a conceptual commitment to use grounded representations and to follow the principles of embodied and situated cognition (see Franklin, Strain, et al., 2013, sec. 4.3). This manuscript expands on this commitment, making its implementations and ramifications more precise in terms of LIDA’s representations, processes, and modules.
3. LIDA is a biologically inspired cognitive architecture that incorporates and elaborates on numerous psychological theories, including many important aspects of the Global Workspace Theory of consciousness (GWT; Baars, 1988). LIDA’s modeling of “functional consciousness” is particularly relevant for this work, as the phenomena of mental simulation, mental imagery (i.e., consciously accessed mental simulations), and simulation-based cognition are best understood with respect to a model of consciousness. While the scientific study of consciousness has become more acceptable in recent years, research into machine consciousness is limited, and the construction of conscious

artifacts (Franklin, 2003) has rarely been attempted. LIDA is one of only a few cognitive architectures that attempt to model consciousness.

4. LIDA attempts to model both human and non-human cognition. Many cognitive architectures are more narrowly focused on human-specific cognition and modeling human-specific activities. As a result, these architectures tend to focus more on declarative memory and propositional, symbolic thought. By contrast, LIDA's sensory, perceptual, and behavior-generating modules can function independently of declarative memory. Consequently, non-symbolic, non-propositional, imagistic processes can be implemented more naturally as extensions of LIDA's existing modules and processes.<sup>4</sup>

---

<sup>4</sup> By comparison, Soar's implementation of mental imagery and non-symbolic reasoning required the addition of several special-purpose, add-on modules (Wintermute, 2012).

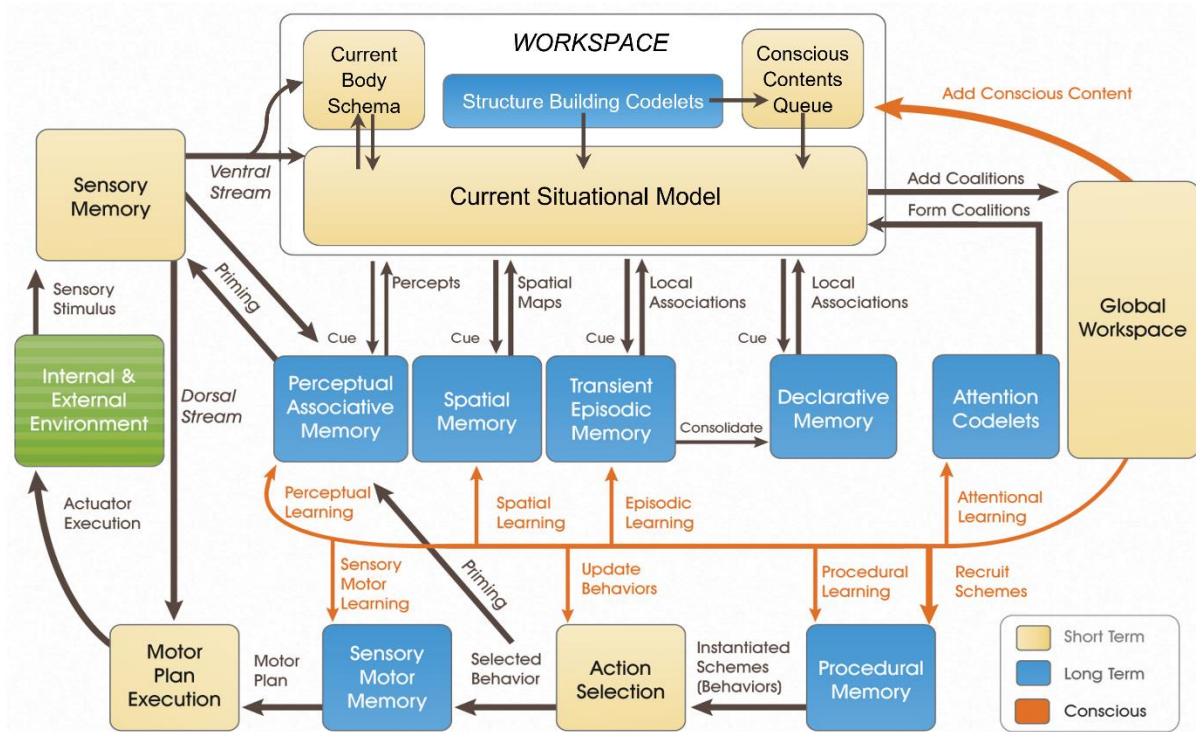


Figure 8. LIDA's cognitive cycle diagram.

### The LIDA Cognitive Cycle

LIDA is composed of many short- and long-term memory modules and supporting cognitive processes (e.g., “codelets,” consolidation, cueing, learning, and decay)—many of these are depicted in Figure 8. All of LIDA's cognitive activities are conceptualized as occurring within, or emerging as the result of, a continual series of potentially overlapping *cognitive cycles*<sup>5</sup>.

LIDA's cognitive cycles are sub-divided into three phases: (1) perception and understanding, (2) attention, and (3) action and learning. Higher-order cognitive processes such as planning, deliberation, and problem solving typically require many cognitive cycles.

<sup>5</sup> The cognitive cycle corresponds to the “action-perception cycle” referred to by many psychologists and neuroscientists (see Dijkstra et al., 1994; Freeman, 2002; Fuster, 2004; Neisser, 1976).

During LIDA's *perception and understanding phase*, sensory stimuli from an agent's environment activate low-level feature detectors in LIDA's Sensory Memory module. These, in turn, activate higher-level perceptual representations in Perceptual Associative Memory (PAM). Perceptual representations receiving sufficient activation can be instantiated as percepts into LIDA's Current Situational Model (CSM)—a sub-module of LIDA's Workspace. Sensory representations, percepts, recalled (cued) long-term memories (e.g., episodes), and the representations generated by structure building codelets (e.g., mental simulations) can co-exist in LIDA's CSM—for example, within its current “Perceptual Scene” (McCall, Snaider, et al., 2010). These preconscious mental representations can influence LIDA's internal dynamics, and they reflect an agent's “understanding” of its current situation.

During LIDA's *attention phase*, attention codelets attempt to identify preconscious representations in the CSM that are of interest to them—based on their individual selection criteria (e.g., novelty, surprise, or urgency). If found, attention codelets will bring that content to a coalition forming process, which may create one or more *coalitions* that include that content. These coalitions then compete in a winner-take-all competition in LIDA's Global Workspace, and the content of the winning coalition is globally broadcast to all of LIDA's modules. An agent is said to be “conscious” of the mental representations in its global broadcasts.<sup>6</sup>

During LIDA's *action and learning phase*, content from its global broadcast is received by all of its modules. Procedural Memory uses that “conscious content” to activate situationally

---

<sup>6</sup> LIDA currently makes no claims regarding phenomenal consciousness. Rather, LIDA attempts to model the functional aspects of consciousness without reference to qualia (i.e., what it is like to experience conscious content). This notion of consciousness is similar to what Block (1995) defines as “access consciousness.”

relevant *schemes*. Schemes are mental representations that encode observed correlations between situational contexts, actions, and the predicted results of those actions. Each scheme additionally has a base-level activation that estimates the likelihood that the scheme's action will produce its expected result (given a similar situational context). Schemes receiving sufficient activation (following a conscious broadcast) are instantiated as *behaviors*. Procedural Memory then sends these behaviors to LIDA's Action Selection module.

Action Selection chooses at most one behavior from its set of selectable behaviors per cognitive cycle.<sup>7</sup> This chosen behavior is referred to as the “selected behavior” for that cycle. Action Selection then sends this behavior to LIDA's Sensory Motor System (SMS; see Dong & Franklin, 2015) for execution. The SMS is composed of two modules: Sensory Motor Memory (SMM) and Motor Plan Execution (MPE). SMM is a long-term memory module that instantiates motor plan templates into *motor plans* based on the current selected behavior. MPE then executes those motor plans through a process of situated, online control. During this online control process, motor commands are sent to an agent's actuators in response to its immediate “situated” concerns.

LIDA's numerous learning mechanisms (see Kugele & Franklin, 2021) can also be invoked during its action and learning phase—as a direct result of a conscious broadcast. These mechanisms support the learning of new representations and the reinforcement of previously learned representations. The former is referred to as “instructionist learning” and the latter as “selectionist learning” (see Edelman, 1987).

---

<sup>7</sup> Selectable behaviors may include non-decayed behaviors from a previous cognitive cycle in addition to recently instantiated schemes (behaviors) from the current cycle.



For a more comprehensive introduction to LIDA, see Franklin et al. (2016). A summary of LIDA's long- and short-term memory modules and codelets is given in Table 6 of the Appendix.

### **Modes of Action Selection**

LIDA supports four modes of action selection: consciously mediated, volitional, automatized, and alarms. *Consciously mediated action selection* is a non-deliberative, relatively reactive form of action selection that requires little cognitive effort, and it typically occurs without an agent's awareness of the intention to pursue some goal (Maes, 1989). In contrast, *volitional action selection* is often deliberative and effortful, and agents are at least partly conscious of the action selection process. During volitional action selection, one or more "options" are considered (usually over many cognitive cycles). If Action Selection chooses one of these options for execution, the agent will typically become consciously aware of that option's goal (i.e., the intention behind that option to action). *Automatized action selection* only occurs when an agent is highly skilled at some activity. It requires that the results of its actions are very predictable, such that they can follow, one after another, with little or no conscious oversight. Finally, *alarms*, refer to the unconscious selection and execution of "urgent" actions. This mode of action selection occurs under exceptional circumstances that require extremely rapid responses (for example, reactively turning the wheel of a car or "slamming on the brakes" while driving to avoid hitting a car that has suddenly cut in front of you).

## The Conscious Learning Hypothesis

LIDA adheres to the Conscious Learning Hypothesis of Global Workspace Theory (GWT; Baars, 1988), which states that all “significant learning” requires consciousness.<sup>8</sup> LIDA and GWT further contend that learning occurs as a direct result of a conscious broadcast, though learning may not occur with every conscious broadcast. This causal relationship between conscious broadcasts and learning does not preclude the possibility of a substantial delay between the onset of a conscious broadcast and subsequent learning based on that conscious content. Such delays occur, for example, during offline consolidation, which supports declarative learning.

## Activation

Activation-related parameters are pervasive properties of LIDA’s mental representations and its codelets (see Kugele & Franklin, 2020a). LIDA has historically classified its activations as either base-level activations, current activations, or total activations. *Base-level activation*<sup>9</sup> is used to describe parameters with relatively slow decay rates that are reinforced based on content in the global broadcast. Base-level activations typically represent learned, historical measures of frequency, recency, and/or utility. *Current activation* refers to parameters with relatively rapid decay rates that generally reflect transitory, module-specific notions of situational relevance. And *total activation*, or simply *activation*, is used to describe all other activation parameters.

---

<sup>8</sup> For example, while LIDA acknowledges that priming effects may occur unconsciously, they are not considered “significant learning” because they are limited in scope and duration.

<sup>9</sup> LIDA’s base-level activation is roughly (conceptually) analogous to ACT-R’s (Anderson et al., 2004) concept of base-level activation, but its meaning is far more varied and module specific. It also has a very different activation source, which is based on LIDA’s conscious broadcasts.

Total activation is often calculated using activation functions that combination base-level and current activations.

### **Motivations in LIDA**

LIDA's motivational system (see McCall et al., 2020) is grounded in *feeling nodes*—PAM nodes with a special parameter called *affective valence*. Affective valence quantifies an agent's immediate hedonic response to events (e.g., eating ice cream), indicating liking (positive affective valence) or disliking (negative affective valence). The magnitude of the affective valence is conveyed by the total activation of a feeling node, and its direction (like or dislike) is conveyed by its valence sign (either + or -). Feeling nodes can be activated either by feature detectors in Sensory Memory (e.g., the presence of sugars in foods) or by cueing from the CSM (e.g., recognizing that you have a straight flush in poker). All LIDA agents must be endowed with a set of built-in feeling nodes that serve as the basis for their value systems.

Kringelbach and Berridge (2009) suggested that “liking” and “disliking” should be distinguished from “wanting” and “dreading,” and that they are implemented in brains using distinct neural pathways. Liking/disliking indicates an immediate hedonic response to an event, while wanting/dreading is an attractive or repulsive force associated with an event. Wanting and dreading are quantified in LIDA by an additional parameter called *incentive salience*, which is further divided into a base-level and current incentive saliences. Base-level incentive salience is a context-invariant attraction or repulsion to an event, which is learned from repeated exposure to that event. Current incentive salience is a context-sensitive attraction or repulsion associated with an event that is modulated by an agent's current situation. Current incentive salience is

based on learned, incentive salience links (as opposed to standard links that only propagate activation).

Motivational learning (see Kugele & Franklin, 2021) involves updating base-level incentive saliences and creating incentive salience links. These learned motivational constructs can direct a LIDA agent's attention, influence action selection, and modulate learning. Base-level incentive salience is updated (by PAM) when a conscious broadcast contains an event with one or more associated feeling nodes. For example, suppose that our agent is eating vanilla ice cream. The agent's tongue receives incoming sensory stimuli corresponding to the ice cream. Based on these sensory stimuli, Sensory Memory may activate a "sweet" (interpretative) feeling node in PAM, which PAM may then instantiate into the CSM. Structure building codelets (SBCs) monitoring the CSM may create an event node structure for this "eating ice cream" event and an "activation" link between the instantiated "sweet" feeling node and the "eating ice cream" event node. If this event node structure (which is now augmented with an activated "sweet" feeling node) comes to consciousness, PAM will receive it and update the base-level incentive salience associated with the "eating ice cream" event node. In this case, only one feeling node is associated with the event; however, in general, base-level incentive salience updates will be a function of the affective valences of all feeling nodes associated with an event.

Current incentive salience, which quantifies the context-sensitive attraction or repulsion associated with an event in a given situation, is transmitted from feeling nodes to event nodes over incentive salience links (McCall et al., 2020, p. 17). The current incentive salience ( $i_c$ ) contributed by a feeling node to a linked event (over an incentive salience link) is defined as its incentive-salience-link-weighted affective valence.

A feeling node’s affective valence is given by  $v = a_t \text{sgn}(v)$ , where  $a_t$  is the total activation of the feeling node at time  $t$  and  $\text{sgn}(v)$  denotes its valence sign (i.e., the “direction” of like or dislike). Therefore, current incentive salience is defined as  $i_c = vw = a_t \text{sgn}(v)w$ , where  $w$  is the weight of a given incentive salience link and

$$\text{sgn}(v) = \begin{cases} -1 & \text{if } v < 0, \\ 0 & \text{if } v = 0, \\ 1 & \text{if } v > 0. \end{cases}$$

A structure building codelet (SBC) can create an incentive salience link between a feeling node and an event node if changes in that feeling node’s activation<sup>10</sup> are attributed to the occurrence of that event. For example, an SBC may notice that the event of “drinking water” is correlated with a decrease in the activation of an agent’s “thirst” feeling node (which has a negative valence sign). In this case, the SBC will create a *positively* weighted incentive salience link between the “thirst” feeling node and the event of “drinking water.” If, on the other hand, an SBC notices that an event, say “exercising,” correlates with an increase in the activation of an agent’s “thirst” feeling node, then the SBC may create a *negatively* weighted incentive salience link between the “thirst” feeling node and the event of “exercising.”

In general, for any incentive salience link between a feeling node and an event node, its weight  $w$  will be given by  $w = \text{sgn}(v)f(\Delta a)$ , where  $\Delta a = a_{t+1} - a_t$  is the change in that feeling node’s activation between conscious broadcasts at time  $t$  and  $t + 1$ ,  $\text{sgn}(v)$  is the feeling node’s valence sign, and  $f: \mathbb{R} \rightarrow [0,1]$  is a function that scales  $\Delta a$  (for example, a sigmoid

---

<sup>10</sup> Conscious broadcasts are stored temporarily in LIDA’s Conscious Contents Queue (CCQ)—a data structure in the preconscious workspace that is accessible to structure building codelets. SBCs can detect changes in the activation of feeling nodes by comparing their activations between subsequent broadcasts in the CCQ.

function). Intuitively, the weights will be positive for events that lead to beneficial changes in an agent’s homeostatic state and negative if they lead to detrimental changes. When a node structure containing a new incentive salience link is consciously broadcast, the new incentive salience link can be learned into PAM and later influence the *current* incentive salience associated with the linked to event node.

### **The Nature of LIDA’s Representations**

LIDA is committed to the principles of embodied and situated cognition, which is interpreted as a “structural coupling” between autonomous agents and their environments (Franklin, Strain, et al., 2013). Grounded representations have also been advocated for throughout LIDA’s development (Agrawal et al., 2018; Franklin, Madl, et al., 2013; Kugele & Franklin, 2020b; Ramamurthy et al., 2006). Apart from these commitments, LIDA is representation agnostic, and individual implementations are free to explore different representational options.

Historically, LIDA’s primary knowledge representation has been *node structures*—directed graphs containing one or more *nodes* and zero or more *links*. Nodes are symbolic representations that refer to objects, entities, events, concepts (etc.), and links indicate semantic relationships between nodes (e.g., “X is a Y,” “X has a Y,” and “X caused Y”). Activation is said to propagate over links, from source nodes to sink (or target) nodes. While nodes structures are, themselves, symbolic representations (similar to semantic networks; see Sowa, 1991/2014), it has been proposed that these data structure could be “grounded” by associations with sensory content originating from LIDA’s Sensory Memory module.

Node structures are conceptually appealing (particularly from an analytical modeling or pedagogical perspective) because they are easily visualized and interpreted by humans.

Unfortunately, they suffer from some computational limitations that have inspired exploration into more scalable representational options (e.g., modular composite representation vectors; see Snaider & Franklin, 2014a). For the purposes of this manuscript, I will generally use node structures for illustrating conceptual details; however, other representations may be used, as necessary, for computational purposes.

### **The LIDA Conceptual Model and Its Implementations**

LIDA's conceptual model (Franklin et al., 2016) specifies a set of high-level components (e.g., modules and processes) and their interactions. And LIDA's conceptual commitments (Franklin, Strain, et al., 2013) constrain how these components *must* operate. LIDA's computational instantiations (i.e., LIDA agents) are required to abide by both LIDA's conceptual model and its commitments; however, within those confines, there is a considerable amount of flexibility related to implementation details, and some of these details are necessarily unspecified by the LIDA model. I say "necessarily unspecified" for at least two reasons. First, an agent may be situated in any of an enormous variety of environments, each with its own character and relative complexity. This environmental sensitivity will likely flavor most, if not all, module implementations. This is a direct consequence of embodiment and situated principles: *all aspects of cognition are shaped by the agent's body and its situated relationship with its environment*. Second, some details are necessarily unspecified due to our imperfect knowledge of minds. Biological minds provide the only known examples that "generally intelligent" autonomous agents are constructible; however, our knowledge of how they accomplish this feat is far from complete.

The LIDA conceptual model and its commitments are a testament to what we believe we know about minds based on our current interpretation of the available evidence; however, the computational implementations of those conceptual constructs are subject to change without invalidating the LIDA conceptual model. Therefore, any computational implementation of a module should be considered *an* implementation not *the* implementation of a module, and it is entirely consistent (and eminently useful) to admit *many* implementations, so long as they are consistent with LIDA's conceptual foundations.

Finally, note that not every LIDA module and process must be implemented in a LIDA agent for it to be considered a LIDA agent. Depending on the nature of the agent one is trying to implement (e.g., reactive, deliberative, etc.) or the research question one is trying to explore (e.g., replicating hippocampal lesion studies), it may be beneficial to focus on a subset of modules or processes, so long as the resulting system remains an autonomous agent consistent with LIDA and its conceptual commitments.



## Chapter 5

### Grounded Representations, Mental Simulations, and Multimodal Perception

It is unlikely that grounded cognition will be fully accepted until classic research paradigms can be understood within its framework. In cognitive psychology, for example, how would a classic paradigm such as recognition memory be understood as grounded? (Barsalou, 2008, p. 635)

This is the first of three chapters detailing LIDA-based implementations for ES-Hybrid’s fundamental components. This chapter focuses on sensorimotor representations, grounded concepts, and the generative processes that support mental simulation (see Chapter 3). A conceptual model and computational approach to multimodal perception and perceptual learning is then developed based on this foundation.

The computational implementation I develop here can be described as a *neuro-symbolic* system.<sup>1</sup> It combines generative artificial neural networks (see Chapter 2) with symbolic data structures (i.e., mental representations). Specifically, I combine beta-variational autoencoders ( $\beta$ -VAEs) with a content-addressable activation graph (containing both modal and amodal nodes). These components and their functions within LIDA will be described in detail below.

#### Background: Variational Autoencoders (VAEs)

Recall from Chapter 2 that autoencoders are generative artificial neural networks that learn in an unsupervised fashion (i.e., from *unlabeled* training data). They combine an *encoder* (or

---

<sup>1</sup> Kautz (2022) referred to this type of neuro-symbolic architecture as **Neuro | Symbolic**.

recognition) network, a *decoder* (or generative) network, and a loss function that is based, in part, on *reconstruction error*. The encoder network transforms its inputs into (typically) lower-dimensional representations—often called latent representations—and the decoder network attempts to *reconstruct* the encoder’s inputs from those latent representations.

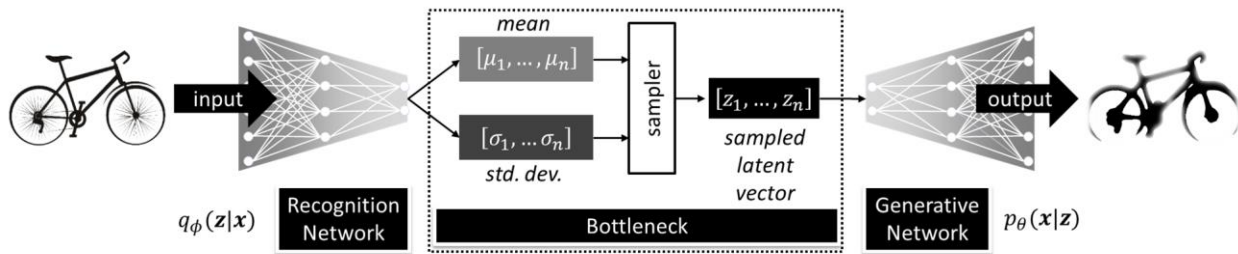


Figure 9. Depiction of a variational autoencoder (VAE). (Note that  $\mathbf{x}$  and  $\mathbf{z}$  are vectors.)

*Variational autoencoders* (VAEs; Kingma & Welling, 2013) learn latent representations that are interpretable as multi-dimensional *probability distributions* (see Figure 9). Specifically, for each input, a VAE’s recognition network outputs a vector of means ( $\boldsymbol{\mu}$ ) and standard deviations ( $\boldsymbol{\sigma}$ ) that characterizes that input’s *generative features* (e.g., its size, rotation, brightness, location, and so on).<sup>2</sup> A stochastic *sampler* can then be used to sample individual latent vectors from these probability distributions.

*Sampled latent vectors* ( $\mathbf{z}$ ) correspond to single points in a VAE’s learned latent (vector) space—that is, they function like the outputs from a traditional autoencoder’s recognition

---

<sup>2</sup> In general, the generative features learned by a VAE are not so easily interpreted by humans. This limitation has been partially addressed in more recent VAE architectures, such as  $\beta$ -VAEs (Higgins et al., 2017), which will be discussed later in this section.

network. These sampled latent vectors can then be fed into the VAE's generative network to create reconstructions of the encoder's inputs. They can also be directly compared using a similarity measure (e.g., cosine similarity). Crucially, vectors that are "close" in a latent space will typically have similar generative features; consequently, the inputs  $\mathbf{x}$  that generated those latent vectors will often *resemble* one another (as will their reconstructions).

One of the benefits of using variational autoencoders over traditional autoencoders is that they tend to produce "smoother" latent spaces. Small perturbations in latent vectors typically correspond to similar input mappings (i.e., similar sensory stimuli). They will also tend to produce similar reconstructions (i.e., mental simulations). Furthermore, a VAE's latent space typically has fewer "gaps." That is, they contain fewer regions with incoherent reconstructions or that lack useful input mappings. This follows from the fact that VAEs map their inputs to a *distribution of points* in their latent spaces rather than a single point. As a result, their latent spaces tend to be more "filled in."

*Beta-variational autoencoders* ( $\beta$ -VAEs; Higgins et al., 2017) augment the standard VAE loss function with a regularizing coefficient ( $\beta$ ) that encourages their networks to learn "disentangled latent representations" (Higgins et al., 2018). Higgins et al. (2018) defined a vector representation as being *disentangled* if it can be "decomposed into a number of subspaces, each one of which is compatible with, and can be transformed independently by a unique symmetry transformation" (Higgins et al., 2018, p. 2). Symmetry transformations are vector operations that selectively modify individual generative features while preserving others; therefore, disentangled latent representations are desirable because they contain generative features (e.g., brightness,

position, and size) that can be individually inspected and manipulated to selectively control aspects of the generative process. Otherwise,  $\beta$ -VAEs are identical to standard VAEs.

The  $\beta$ -VAE loss function ( $\mathcal{L}$ ) appears below:

$$\mathcal{L}(\theta, \phi; \mathbf{x}, \mathbf{z}, \beta) = -\mathbb{E}_{q_{\phi}(\mathbf{z}|\mathbf{x})}[\log(p_{\theta}(\mathbf{x}|\mathbf{z}))] + \beta D_{KL}(q_{\phi}(\mathbf{z}|\mathbf{x}) || p(\mathbf{z})) ,$$

where  $\mathbf{x}$  represents a sample of input data (encoder inputs);  $\mathbf{z}$  represents sampled latent vectors (decoder inputs);  $\beta$  is a configurable parameter that controls the amount of “disentangling pressure”;  $D_{KL}$  the Kullback-Leibler divergence; and  $\theta$  and  $\phi$  correspond to the parameters—weights and biases—associated with the encoder and decoder neural networks, respectively. (The first term in the  $\beta$ -VAE loss function corresponds to the reconstruction error.)

Note that  $\beta = 1$  corresponds to a classical variational autoencoder.  $\beta > 1$  constrains the capacity of the learned latent vectors, encouraging disentangling at the cost of lower-quality reconstructions. For more details on the  $\beta$ -VAE’s loss function and its derivation see Higgins et al. (2017).

## Implementation

This section contains my implementations of foundational ES-Hybrid components in LIDA. I implement the requisite modules, representations, and processes, and then illustrate how they can be combined to support multimodal perception, mental simulation, and grounded concept learning. I assume readers are familiar with LIDA’s cognitive cycle, modules, processes, basic representational formats (i.e., node structures), and basic conceptual commitments (e.g., conscious learning). These were introduced in Chapter 4. LIDA’s modules and codelets are also summarized in Table 6 of the Appendix.

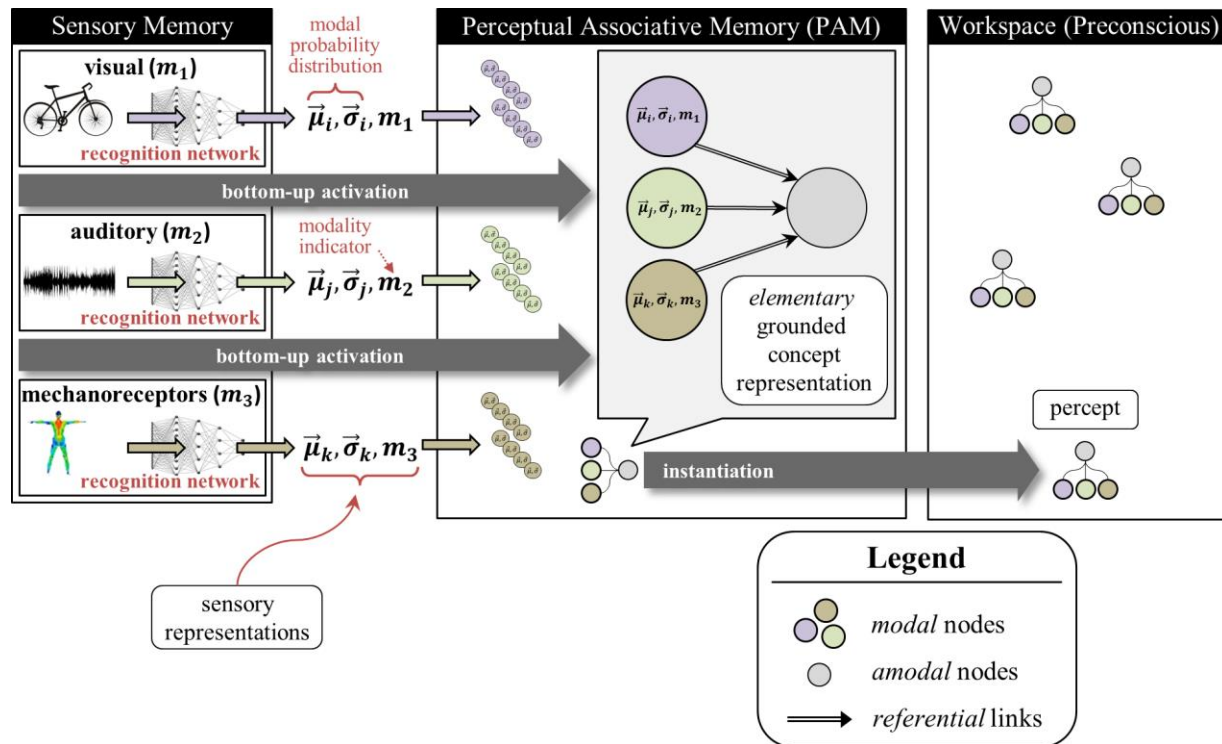


Figure 10. Sensory Memory and bottom-up perception. Modality-specific recognition networks in LIDA’s Sensory Memory module generate modal probability distributions (i.e., sensory representations) that are used to activate *modal* nodes in PAM. Activation can spread through PAM’s activation graph over directed links (e.g., referential associations), activating *amodal* nodes that symbolize grounded concepts. An *elementary* grounded concept (see Chapter 3) is depicted above. Node structures receiving sufficient activation are instantiated into LIDA’s preconscious Workspace as percepts.

### *Sensory Memory and Sensory Representations*

Sensory Memory is a short-term memory module that encodes modality-specific sensory content as patterns of activation over *low-level feature detectors*<sup>3</sup>. Environmental stimuli activate these feature detectors, resulting in the generation of *sensory representations* (cf. ES-Hybrid’s sensorimotor representations; Chapter 3). Like their originating low-level feature detectors, sensory representations are modality-specific—they encode features from a single sensory

<sup>3</sup> The feature detectors learned by ANN architectures, such as convolutional neural networks (see Chapter 2), are typically hierarchically organized. In their earliest layers, they might encode feature maps that are receptive to edges (in various orientations) or colored regions. In their later layers, they might encode feature maps that are receptive to more elaborate shapes (e.g., eyes; see Mahendran & Vedaldi, 2016; Z. Qin et al., 2018.)

modality (e.g., visual, auditory, somatosensory)<sup>4</sup> Sensory representations can be characterized as non-symbolic representations that are *modal*, *analogical*, and *generative* (see Chapter 3 for a review of these properties).

Sensory Memory can be implemented using a *set* of modality-specific  $\beta$ -VAEs—one per sensory modality (see Figure 10).<sup>5</sup> The  $\beta$ -VAEs’ recognition networks implement Sensory Memory’s low-level feature detectors. And the (modal) probability distributions generated by these networks implement LIDA’s sensory representations.

Sensory representations are arguably the most important representations for implementing embodied, stimulation-based (grounded) cognition in LIDA. They directly support the following functions:

- (1) Sensory representations characterize the most “important” features of sensory stimuli; therefore, they can serve as sensory signatures that support the *identification* of, and *discrimination* between, those sensory stimuli when present in an agent’s (internal or external) environment.<sup>6</sup>
- (2) Sensory representations are used to activate perceptual representations in LIDA’s PAM module; thus, they directly support bottom-up perception.

---

<sup>4</sup> LIDA does not require feature detectors or sensory representations to be modality-specific. This is an additional constraint imposed by ES-Hybrid.

<sup>5</sup> The specifics of the  $\beta$ -VAEs’ network architectures (e.g., feed-forward, recurrent, convolutional) will depend on the representational needs of each sensory modality. The flexibility to vary these network architectures by sensory modality is one benefit of the implementation used in this chapter.

<sup>6</sup> Harnad (1990, sec. 3.1) defined *discrimination* as the ability to judge the extent to which two representations are the same or different (i.e., their degree of similarity), and he defined *identification* as the ability to assign a unique (symbolic) identifier (i.e., a “name”) or category label to a class of inputs.

- (3) Sensory representations are sent to LIDA's Current Situational Model (CSM), where they contribute to an agent's understanding of its current situation.
- (4) Sensory representations are sent to LIDA's Sensory Motor System (SMS), where they support *situated, online control processes* (see Chapter 2).
- (5) Sensory representations are the *grounding* constituents of grounded concept representations (see Chapter 3).
- (6) Sensory representations are used by generative processes to construct mental simulations.

Many of these functions will be expanded on below.

### ***Perceptual Associative Memory and Grounded Concepts***

Perceptual Associative Memory (PAM) is LIDA's "recognition memory" (Franklin et al., 2016, sec. 5.2.1). It is a long-term memory module that supports the identification of objects, entities, situations, events, and their properties. It is also LIDA's long-term memory for grounded concepts.

I implement PAM as a *content-addressable activation graph*. It is composed of a set of *nodes* that can be connected using directed *activation links*.<sup>7</sup> It is an *activation graph* because current activation (see Chapter 4) can propagate between its nodes along activation links. And it is a *content-addressable* data structure because its representations are typically activated using resemblance-based comparisons (and spreading activation) rather than name-based lookups.

---

<sup>7</sup> Nodes can also be connected using incentive salience links (see Chapter 4). In this chapter, I omit this and other details of LIDA's motivational system to simplify and focus the exposition.

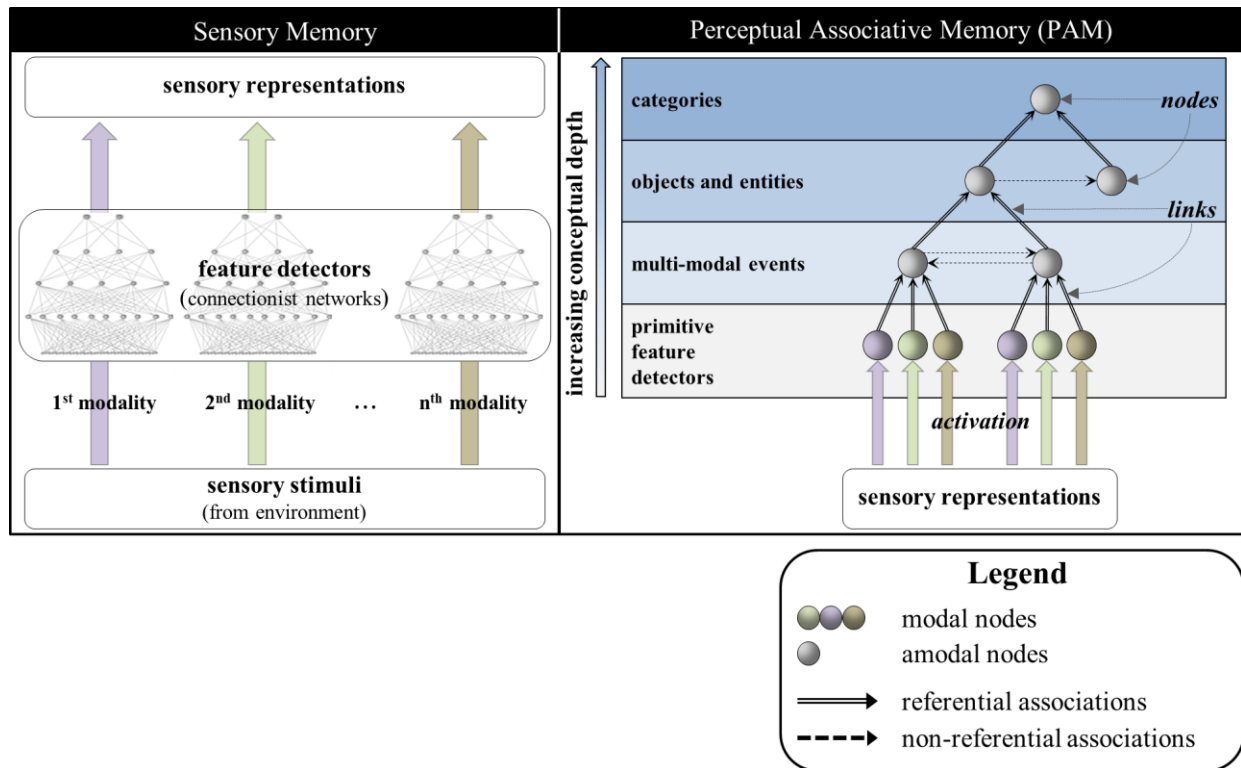


Figure 11. Grounded representations and conceptual generalization.

PAM’s nodes represent grounded concepts and instance of those concepts (e.g., objects, entities, situations, events, features). These can range in specificity from individual experiences to highly generalized category representations. As such, PAM’s activation graph is typically hierarchically structured (see Figure 11, Right Panel).

My implementation subdivides PAM’s nodes into two types: modal and amodal. Modal nodes are modality-specific nodes with associated sensory representations (e.g.,  $\beta$ -VAE generated modal probability distributions) and a modality indicator (e.g., visual, auditory, tactile). Modal nodes receive current activation *exclusively* from Sensory Memory; therefore, they function as PAM’s “primitive feature detectors” (see McCall, Snider, et al., 2010). Sensory Memory activates modal nodes by comparing their sensory representations and those for



incoming environmental stimuli. Modal nodes receive current activation in proportion to the degree of similarity (e.g., cosine similarity) between those sensory representations.

In contrast, amodal nodes *do not* have associated sensory representations or modality indicators. They are *symbolic* representations that receive activation *exclusively* from other PAM nodes: they are never (directly) activated by Sensory Memory. Therefore, amodal nodes function as PAM’s “non-primitive feature detectors” (see McCall, Snaider, et al., 2010). PAM’s amodal nodes never occur in isolation, as such representations would be functionally inert. And amodal nodes ultimately depend on modal representations for activation.

I implement *elementary* grounded concepts (see Chapter 3) in LIDA using a combination of modal and amodal nodes. Specifically, I use a “hub-and-spoke” (Patterson et al., 2007; Ralph et al., 2010, 2017) representational format. Coordinating amodal nodes serve as “hubs” that holistically symbolize distinct, potentially multimodal, experiences. Modal nodes serve as “spokes” that ground those amodal symbols in sensory content. Modal nodes are connected to their amodal hubs using “referential” activation links (described later in this chapter). More elaborate grounded concept representations (e.g., for objects and categories) can be created by binding these elementary grounded concepts together into more generalized and complex structures—supported by “referential” activation links and hub-and-spoke-style topologies (see Figure 11, Right Panel).

All PAM nodes have a *current activation* (representing its current situational relevance) and a *base-level activation* (representing its historical frequency, recency, and salience in “conscious” broadcasts). A node’s *total activation* is a function (e.g., summation) of these parameters. PAM nodes with sufficient total activation are instantiated into LIDA’s Current

Situational Model (CSM) as percepts (see Figure 10). Note that all activations decay over time, with current activations typically decaying much more rapidly than base-level activations (see Chapter 4, Activation).

### ***Referential and Non-Referential Associations***

Recall from Chapter 3 that ES-Hybrid specifies two kinds of representational associations: referential and non-referential. Referential associations establish a correspondence between two things. For example, they can specify constitutive (*part-of*), identity (*is-a*), and membership (*kind-of*) relationships. Referential associations are *grounding* associations. Non-referential associations, on the other hand, characterize non-correspondence-based relationships. These include causality, co-occurrence, temporal ordering, spatial relationships, etc. Non-referential associations are non-grounding associations.

Referential associations hold a “privileged” status with respect to ES-Hybrid’s conception of grounded cognition and this LIDA-based implementation, as they are used to ground its conceptual representations.<sup>8</sup> As such, referential associations depend, at least in part, on innate (built-in) cognitive processes for their creation and interpretation. These processes might oversee multimodal sensory binding, conceptual generalization, and the generative processes that govern mental simulation (among others). In contrast, non-referential associations may largely depend on learned conceptual relationships that are themselves grounded concepts.

---

<sup>8</sup> Referential associations are *necessary* for grounding concept representations, but they are not *sufficient*. Referential associations can connect ungrounded concepts.

Referential and non-referential associations can be implemented in LIDA as types of activation links.<sup>9</sup> Like standard activation links, referential and non-referential activation links are created by structure building codelets (see Chapter 4), and they can be learned into long-term memory if they are included in a global broadcast. In the current implementation, the distinction between these two types is primarily of concern to structure building codelets (SBCs; see Chapter 4), for example, “simulator” SBCs, which will be discussed later in this chapter. Specifically, SBCs create these associations and may interpret the representations containing them differently depending on the link type (e.g., as grounded vs ungrounded concepts). Link types could also differentially affect the activation dynamics in long-term memory modules, such as PAM, though I will not elaborate on this idea further here.<sup>10</sup>

The idea of “typed” activation links has precedent in McCall, Franklin, and Friedlander’s (2010) proposal to add what they called “primitive link classes” to LIDA. Primitive link classes were intended to serve as semantic labels (cf. semantic networks; Sowa, 1991/2014) that characterized the associative relationships denoted by activation links. Their proposal included many types of link classes, including those for features (*is-a-feature-of*), parts (*is-a-part-of*), spatial relationships (e.g., *is-left-of* or *is-above*), causal relationships (e.g., *is-caused-by*), thematic roles (e.g., *is-an-agent*, *is-a-location*, *is-an-object*), and category membership (e.g., *is-a-kind-of*). As such, referential and non-referential links could be conceptualized as higher-order categories of link classes that group these classes into more basic types.

---

<sup>9</sup> Referential and non-referential associations *do not* apply to LIDA’s incentive salience links. Incentive salience links specify the current motivational significance of an object, entity, or event—not what it is, or how it relates to other things in an environment.

<sup>10</sup> See link “types” and “labels” in Hofstadter and Mitchell’s (1994) Copycat architecture for ideas on how this could be accomplished in LIDA.

### *Simulator Structure Building Codelets*

Barsalou defined a “simulator” as the knowledge and generative processes that support concept representation and mental simulation (Barsalou, 1999, sec. 2.4.3). That is, in Barsalou’s theory, simulators are both mental representations *and* generative processes. My implementation separates these concerns. Specifically, my implementation uses “simulator” structure building codelets (SBCs)—generative processes that construct mental simulations for grounded concept representations. Simulator SBCs are not equated with individual concepts; they merely support their mental simulation. And a single simulator SBC can support the simulation of many grounded concepts.

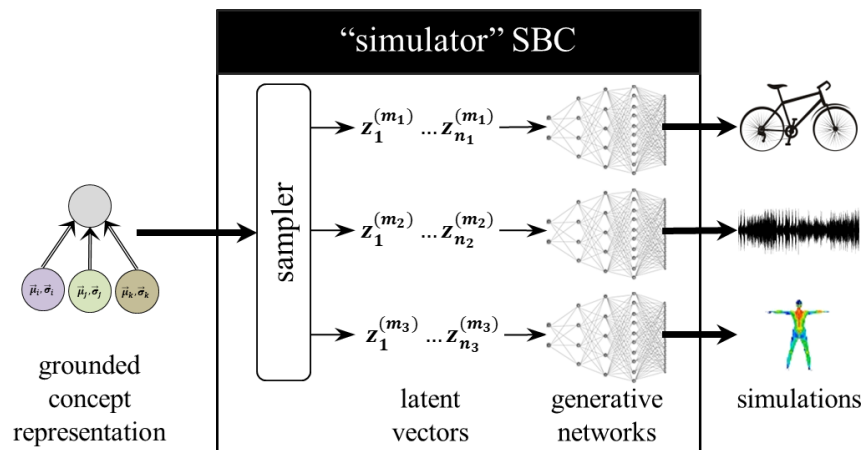


Figure 12. Simulator structure building codelet (SBC).

Conceptually, simulator SBCs *re-activate* portions of LIDA’s Sensory Memory. They do this in a *top-down* fashion, proceeding from grounded concepts to their grounding sensory representations over a chain of *referential* links. Simulator SBCs generate modality-specific

mental simulations for those sensory representations and integrate them into multimodal structures in LIDA's Current Situational Model (CSM).<sup>11</sup>

Computationally, simulator SBCs can be implemented using a set of modality-specific  $\beta$ -VAE *generative networks* and a stochastic *sampler* (see Figure 12). Mental simulation involves first sampling latent vectors for each sensory representation associated with a grounded concept and then feeding those sampled latent vectors into the  $\beta$ -VAEs' generative networks. The outputs from these generative networks are modality-specific, sensory-like representations—that is, internal reconstructions of (internal or external) environmental stimuli. These modality-specific simulations can then be integrated into more complex, multimodal sensory scenes that are accessible from LIDA's CSM. Portions of these sensory scenes may cue relevant percepts (from Perceptual Associative Memory) or be operated on by structure building codelets. This sensory content might also be included in LIDA's global broadcasts, if selected by one or more attention codelets.

In summary, simulator SBCs construct mental simulations by iteratively re-activating portions of Sensory Memory. Mediating amodal nodes and referential links provide a bridge from grounded concepts to their grounding sensory representations. Once constructed, simulator SBCs associate their mental simulations with their originating grounding concept representations in the preconscious workspace. The resulting combination of “virtual” sensory content and grounded node structures is reflected in LIDA's Perceptual Scene (see Chapter 7).

---

<sup>11</sup> More precisely, mentally simulations are added to the sensory portion of the CSM's Perceptual Scene (McCall, Snaider, et al., 2010), and their originating grounded concepts are added to its node layer. Grounded concepts and their mental simulations are then linked together via referential associations. The simulation of more complex, multi-part concept instances requires a process of *sequential elaboration* that works in concert with an agent's introspective processes. These details will be discussed in Chapter 7.

### *Simulation-Based Attention*

Barsalou (1999, sec. 2.4.3) characterized learning as the establishment of simulators. Moreover, he equated simulators with concepts and one’s understanding of those concepts with the ability to simulate them. These hypotheses, if correct, lead to two interesting corollaries. First, conceptual understanding can be *operationalized* by one’s overt and covert generative capabilities. Second, conceptual learning can be *optimized* by biasing one’s attention towards objects, entities, and events for which one’s mental simulations are inadequate. That is, generative deficits can serve as cues that more experiential learning is needed (with respect to a concept).

These observations suggest an attentional process. For example, the  $\beta$ -VAE’s reconstruction error,

$$\mathcal{E}(\theta, \phi; \mathbf{x}, \mathbf{z}) = -\mathbb{E}_{q_{\phi}(\mathbf{z}|\mathbf{x})}[\log(p_{\theta}(\mathbf{x}|\mathbf{z}))] ,$$

could be used to guide an agent’s attention towards salient environmental stimuli. In particular, a high reconstruction error could be considered a measure of “surprise.”<sup>12</sup>

This reconstruction-error-based attentional process is primarily useful during *bottom-up perception* since reconstruction errors can only be determined in the presence of a concurrent, veridical, sensory signal—error calculation requires a basis for comparison. This suggests that simulation-based autonomous agents *continually* attempt to simulate their sensory experiences, which may seem counter-intuitive. Why construct “virtual” sensory content when the “real”

---

<sup>12</sup> A LIDA agent could become consciously aware of those “surprising” stimuli if reconstruction error were used as an activation source for a “surprise” feeling node.

sensory content (environmental stimuli) is readily available? And, if mental simulation is continually occurring, how do mental simulations so frequently escape our notice?

This idea of continual mental simulation becomes less absurd when we make the connection between mental simulation and prediction. Mental simulation *is* a form of experience-based prediction, and Moulton and Kosslyn (2009) argued that this predictive function of mental imagery is its *primary* function.

Albright (2012) referred to the idea that non-volitional mental imagery (see Chapter 7) can influence and support perception as *the implicit (or automatic) imagery hypothesis*. This hypothesis states that one's perceptual experiences typically depend on both "real" stimuli and "virtual" stimuli (i.e., mental simulations). The degree to which perception relies on one or the other depends on the quality of the "real" sensory stimuli, and one's knowledge of their environment (e.g., their ability to generate mental simulations of those concepts). Albright argued that, ordinarily, mental simulations serve "to augment sensory data with 'likely' interpretations [i.e., predictions]" (Albright, 2012, p. 235) in order to compensate for noisy, ambiguous, and partial sensory information. Though controversial, there is a wealth of empirical and theoretical support for the idea that perception is based on a complex interplay between the bottom-up signals resulting from incoming sensory stimuli and internally generated, top-down signals (e.g., resulting from mental simulations; Bar, 2009; Bruner et al., 1951; Cope et al., 2017; N. Dijkstra et al., 2017; Farah, 1985, 1989; Hansen et al., 2006; Mast et al., 2001; Moulton & Kosslyn, 2009; O'Callaghan et al., 2017; Powers III et al., 2016; Tian et al., 2018).

According to this view, mental simulation is fundamental to perception. When there are no discrepancies between our perceptions and corresponding mental simulations, we are

typically unaware that these continual background mental simulations exist. However, when there are discrepancies (incorrect predictions), we may become conscious of them and experience associated feelings of confusion or surprise. These discrepancies may also indicate that an agent’s internal representations, generative processes, and predictive models of the world need updating.

### ***Cognitive “Object” Maps***

Previous LIDA research (e.g., Madl et al., 2018) has explored allocentric<sup>13</sup>, topographically organized, “cognitive maps” (Schiller et al., 2015; Tolman, 1948). These hybrid (symbolic/non-symbolic) data structures encode the spatial locations of objects and places within an agent’s environment, and support agent localization and route planning, among other things. LIDA’s implementation of cognitive maps relies on *place nodes*—PAM nodes that represent distinct locations within a topographic or volumetric extent.<sup>14</sup> Cognitive maps depict the spatial dimensions<sup>15</sup> within an environment (e.g., they may depict the layout of one’s house), and place nodes are overlaid on those depictions—like pins in a map—to symbolize specific locations with that spatial extent.

An object’s location can be specified within a cognitive map by associating a PAM node that symbolizes that object with one or more of its place nodes (see Figure 13). Note that

---

<sup>13</sup> *Allocentric representations* rely on agent-external landmarks and world-centric (or god’s eye) points of view to represent the locations of things in an environment. This contrasts with *egocentric representations* that represent locations relative to an agent’s own position or with respect to their individual points of view.

<sup>14</sup> Place nodes are inspired from “hippocampal place cells” that occur in animal brains, which are believed to encode an animal’s belief about its current spatial location.

<sup>15</sup> Temporal cognitive maps are also possible, though they have yet to be explored in the context of LIDA.



cognitive maps are typically hierarchically organized: a single spatial region may be represented by many cognitive maps (concurrently) with different scales and resolutions.

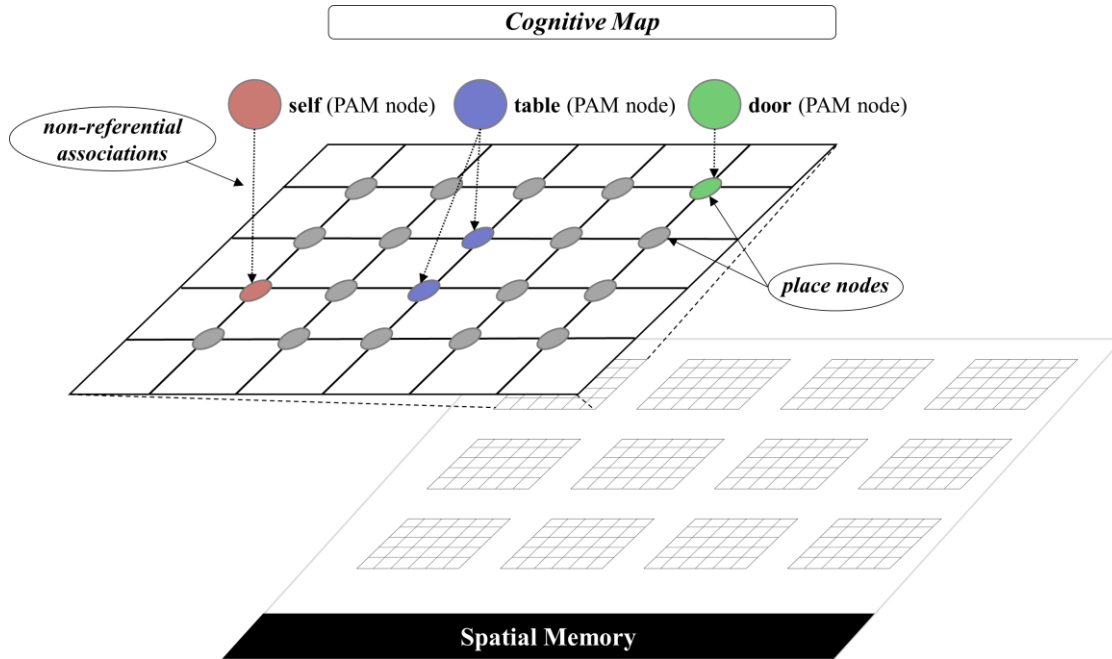


Figure 13. Spatial cognitive map.

Object-centric cognitive maps that encode the relationships between object parts or scene elements (from egocentric or object-centered reference frames) could also be implemented. For example, this type of cognitive map could be retinotopically organized, where locations on the cognitive map correspond to locations in an agent’s visual field. Alternately, they could encode the tactile relationships between object parts, such as would occur when someone explores an object with their hands. Or they could be used to coordinate an agent’s somatosensory (tactile, temperature, proprioceptive, nociceptive) inputs with respect to its body’s extent (e.g., a somatosensory homunculus<sup>16</sup>). These cognitive “object” maps are similar to Kosslyn’s “object

---

<sup>16</sup> Somatosensory cognitive (or perceptual) maps such as these could be used to implement LIDA’s current body schema (see Neemeh et al., 2021).

maps” (Kosslyn, 1994; Kosslyn et al., 2006) or simplified versions of Barsalou’s modal “frames” (Barsalou, 1999, fig. 3).

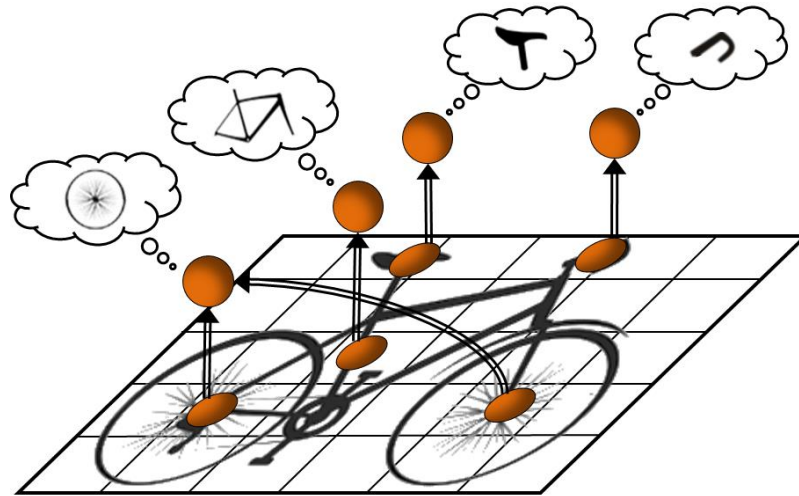


Figure 14. A schematic diagram depicting a cognitive “object” map.

Cognitive “object” maps could be implemented in LIDA by combining PAM’s node structures with topographic, retinotopic, or volumetric extents that occur within its various short-term (e.g., Sensory Memory and the Current Situational Model) and long-term (e.g., Spatial Memory) memory modules. For example, the bicycle in Figure 14 combines a topographic extent corresponding to a bicycle with a set of node structures that describe the details and locations of each of its part (i.e., sub-regions of interest). These descriptive node structures are anchored to the bicycle’s extent using object-centric “place nodes” that specify distinct locations along that extent. Like allocentric cognitive maps, cognitive “object” maps can be hierarchically organized, concurrently supporting multiple scales and resolutions.

### ***Multimodal Perception***

Perception involves the recognition (identification) of objects, entities, situations, and events within environmental stimuli. During perception, these stimuli are mapped to learned or innate

mental representations (e.g., grounded concepts) that provide the most plausible explanations for those stimuli, and the resulting “percepts” are integrated into an agent’s internal model of its current situation (i.e., its Current Situational Model).

In practice, perception is an active process, supported by predictions, exploration, and speculative reasoning. And it can be viewed as the result of both “bottom-up” and “top-down” processing (e.g., see N. Dijkstra et al., 2017; Intaité et al., 2013; Mechelli et al., 2004). *Bottom-up processing* is characterized by the *forward* flow of information. With respect to LIDA, this begins with the arrival of sensory stimuli in Sensory Memory and ends with a global (conscious) broadcast. *Top-down perceptual processing*, on the other hand, is characterized by the *backward* (or recurrent) flow of information and internally generated signals. This can take the form of mental simulations, predictive processing (Clark, 2013), and situational expectations.<sup>17</sup> With respect to LIDA, top-down processing is reflecting in the dynamics of LIDA’s (preconscious) Workspace, including the activity of its structure building codelets (e.g., simulator SBCs) and the cueing of long-term memory modules (e.g., Declarative Memory and Perceptual Associative Memory).

In this section, I will primarily focus on developing an account of LIDA’s bottom-up perceptual processing. This partial account of LIDA’s perceptual processes is based on the module, representation, and process implementations described earlier in this Chapter. Chapters 6 and 7 delve into more detail about top-down cognitive processes.

---

<sup>17</sup> Recall from Chapter 3 that prediction and mental simulation are not necessarily distinct mental phenomena. Moulton and Kosslyn (2009) posited that mental simulation is primarily used to create experiential predictions, and numerous researchers have noted a deep connection between prediction, perception, and mental simulation (Barsalou, 2009; Clark, 2013; Jeannerod, 2001). Though I did not mention them explicitly earlier, expectations are clearly a related idea.

LIDA's bottom-up perceptual processing is partially depicted in Figure 10; it involves the following steps:

- (1) Perception begins when environmental stimuli activate Sensory Memory's modality-specific, low-level feature detectors (i.e.,  $\beta$ -VAE recognition networks). Their patterns of activation result in the generation of a set of modality-specific sensory representations (i.e., modal probability distributions).
- (2) Sensory Memory updates the *current activations* associated with Perceptual Associative Memory (PAM)'s modal nodes (i.e., PAM's primitive feature detectors). This is accomplished by a resemblance-based comparison between the sensory representations generated in step (1) and the learned (or built-in) sensory representations grounding PAM's modal nodes. The greater the similarity between these sensory representations, the greater the increase in their modal node's current activation. For computational purposes, current activation is calculated as a function of the *cosine similarity* between two sensory representations.
- (3) *In parallel* to step (2), the sensory representations generated in step (1) are integrated into LIDA's Current Situational Model. *Multimodal-binding* structure building codelets construct elementary grounded concept representation from each such set of co-occurring, modality-specific, sensory representations. These new elementary grounded concepts characterize an agent's individual sensory experiences.
- (4) Current activations propagate in Perceptual Associative Memory (PAM) activation graph over referential links, from modal nodes (i.e., primitive feature detectors) to their connected grounded concepts (i.e., non-primitive feature detectors). When current

activation propagates from multiple (source) nodes to a single (sink) node, their current activations can combine to jointly activate that target node<sup>18</sup>. A portion of this current activation can then propagate to other perceptual/conceptual representations over activation links.

Base-level and current activations combine in PAM to determine a node's total activation. PAM node structures with total activations over an *instantiation threshold* are instantiated into LIDA's preconscious Workspace—i.e., its Current Situational Model—as percepts.

- (5) A simulator structure building codelet (SBC) continually scans the Current Situational Model looking for percepts without associated mental simulations. For each such percept, the simulator SBC constructs a modal simulation. It then associates this modal simulation with its corresponding percept in LIDA's Perceptual Scene (using a referential association). This “virtual” (internally sourced) sensory content comingles with “real” (externally sourced) sensory content within the Perceptual Scene (see Chapter 7).
- (6) Attention codelets scan the preconscious representations in LIDA's Current Situational Model and select among these based on their own interests (i.e., their matching criteria). Selected representations are sent to a coalition forming process, which constructs coalitions from them and sends them to the Global Workspace.

---

<sup>18</sup> When a PAM node is activated from multiple sources, those source nodes function like a heterogeneous, multimodal “stacked ensemble” (Naimi & Balzer, 2018; Sesmero et al., 2015; D. H. Wolpert, 1992). That is, the current activation of the target (sink) node is calculated as a weighted combination of multiple sources of resemblance-based “evidence.” Such a multimodal ensemble of nodes could be used to provide a more robust predictor of a node's situational relevance.

(7) The Global Workspace conducts an activation-based, winner-take-all competition among its coalitions and globally broadcasts the winning coalition's content.

### ***Perceptual (Conceptual) Learning***

LIDA's perceptual (conceptual) learning can be characterized as both instructionist and selectionist (see Edelman, 1987). Instructionist learning involves the learning of new knowledge representations (e.g., sensory representations and grounded concepts). Selectionist learning involves updating/fine-tuning parameters associated with existing knowledge representations (e.g., base-level activations).

**Instructionist Perceptual Learning.** A *multimodal-binding* structure building codelet (SBC) can create new elementary grounded concept representation in LIDA's Current Situational Model by binding together co-occurring sensory representations. It first creates a new *modal* node for each unbound sensory representation and assigns a value to its modality indicator. It then binds these modal nodes to an amodal node using referential links. This new amodal node holistically symbolizes that experience. Other elements of the agent's current situation (i.e., background contexts) may also be associated with these sensory experiences using *non-referential links*. If these new elementary grounded concept representations and their associated contexts are attended to by attention codelets, they may be consciously broadcast and learned into Perceptual Associative Memory (PAM).

More elaborate and generalized node structures could be created in an agent's CSM by other structure building codelets. For example, a structure building codelet could specialize in creating cognitive "object" maps (see Figure 14). These could be created using a process similar to the "schematic symbol formation process" described by Barsalou (1999) for learning frames.

This process begins by first constructing (or allocating) a coordinating non-symbolic representation (e.g., a spatial extent) that characterizes an object's overall shape and properties. For example, the spatial extent for an object could be induced into the sensory portions of LIDA's Perceptual Scene by a simulator SBC. This proto-cognitive "object" map could then serve as an object-centered reference frame within which specific sub-regions could be elaborated.<sup>19</sup> In particular, grounded concept representations (e.g., object parts) could be associated with *object-centric place nodes* that specify distinct locations within an object overall extent.<sup>20</sup> These object maps could then be incrementally learned over multiple cognitive cycles.

Finally, a *generalization* structure building codelet could identify and construct categories of objects, entities, or events (etc.) by (1) identifying clusters of related items in the CSM, (2) creating new amodal category nodes (as necessary) for those related items, and (3) linking identified category members to those category nodes (using referential links). Each of these structures could then be learned into PAM, if they are included in a conscious broadcast.

**Selectionist Perceptual Learning.** If PAM receives a conscious broadcast that contains previously learned PAM nodes, it increases the base-level activation of each such PAM node. The magnitude of this increase is based on the "strength" of the conscious broadcast (i.e., the activation of the winning coalition). Additionally, the parameters (weights and biases) associated with Sensory Memory's recognition and generative networks could be updated from the content

---

<sup>19</sup> Kosslyn suggested that the creation of multi-part (visual) mental images may begin with a "global image" (1994, p. 292) whose parts are elaborated on, as needed. The process advocated for here follows a similar idea: incremental elaboration of an initially low fidelity coordinating non-symbolic representation.

<sup>20</sup> The idea of object-centric place nodes was described in the section on Cognitive "Object" Maps earlier in this chapter.

in a conscious broadcast—based on the  $\beta$ -VAE loss function and stochastic gradient descent. This update requires (1) environmental stimuli (i.e., “real” sensory content), (2) their simulations (“virtual” sensory content), and (3) their sensory representations (modal probability distributions). This required content is available in LIDA’s Perceptual Scene (see Chapter 7); therefore, it could be included in a conscious broadcast.

## **Evaluation**

This chapter detailed conceptual and computational implementations for grounded representations, mental simulation, and multimodal perception in LIDA. While these implementations were developed in accordance with the guidelines specified in Chapter 3, its properties should be confirmed by experimentation and analysis. Specifically, ES-Hybrid requires that grounded concept representation satisfy the following four properties:

- (1) They must be *analogical*, bearing a resemblance to the things they signify.
- (2) They must be *generative*, supporting modal mental simulations.
- (3) They must be *grounded*, either directly or indirectly, in sensorimotor (i.e., sensory) representations.
- (4) They must be *perceptual*, capable of activating and being interpreted by perceptual systems.

In this section, I will demonstrate that my computational implementation satisfies these requirements.



## *Analogical*

Analogical representations are non-symbolic representations that bear an iconic relationship with the things they signify (see Peirce’s semiotics, Chapter 2). As such, they can serve as proxies for their referents (e.g., environmental stimuli) in resemblance-based comparisons. Two analogical representations should be judged similar *by an agent’s perceptual processes* if and only if their referents resemble one another *with respect to that agent’s sensory system*.

I contend that the implementations of LIDA’s Sensory Memory and Perceptual Associative Memory (PAM) modules detailed in this chapter can satisfy this analogical requirement. To test this claim, a  $\beta$ -VAE with a convolutional architecture<sup>21</sup> was trained on Fashion MNIST (Xiao et al., 2017)—a well-known data set containing 70,000 grayscale images (28 x 28 pixels each) from ten different categories of “fashion products” (t-shirts, trousers, pullovers, dresses, coats, sandals, shirts, sneakers, bags, and ankle boots). The resulting  $\beta$ -VAE had 538,529 parameters (i.e., weights and biases), its sampled latent vectors ( $\vec{z}$ ) had 64 dimensions, and its  $\beta$  value was 1.2. Training was fully unsupervised (see Chapter 2) and consisted of five training epochs over the data set’s 60,000 *training* images. The data set’s 10,000 *test* images were withheld during training and used exclusively for experimentation.

To test the  $\beta$ -VAE’s ability to capture the resemblance-based characteristics of its inputs (i.e., whether its sensory representations are analogical), 25 images were randomly sampled from each of the dataset’s 10 categories. The resulting 250 images (sensory stimuli) were used to

---

<sup>21</sup> Alternating convolutional, max pooling layers (2 of each) were used to implement the  $\beta$ -VAE’s encoder network. Transpose convolutional layers were used to implement the  $\beta$ -VAE’s decoder network. Rectified-linear (RELU) activation functions were used throughout.

generate sensory representations. Current activations were then calculated for each pair of sensory representations to determine their apparent perceptual similarity. The resulting values were used to generate the heatmap shown in Figure 15 (Left Panel).

This experimental setup was intended to imitate Sensory Memory's activation of modal nodes in PAM from sensory representations for incoming environmental stimuli. Each row of the heatmap in Figure 15 (Left Panel) can be viewed as a pattern of activation induced over PAM's modal nodes given some sensory input (in this case, an image).

Current activations were based on the cosine similarities ( $\delta$ ) calculated over pairs of sensory representations (i.e., modal probability distributions). These cosine similarities were then used as inputs to a sigmoidal current activation ( $\alpha_c$ ) function,<sup>22</sup>

$$\alpha_c(\delta) = \frac{1}{1 + e^{(-15\delta+10)}} .$$

Note that current activations immediately decayed to zero following each calculation; therefore, current activations were unaffected by earlier trials.<sup>23</sup>

---

<sup>22</sup> While not strictly necessary, it was beneficial to scale cosine similarities using this non-linear activation function to reduce noise. A sigmoidal scaling function such as this will be particularly useful in a LIDA agent, where current activation can gradually accumulate over time because current activation does not immediately decay to zero between subsequent environmental stimuli.

<sup>23</sup> As a side note: residual current activation in PAM's activation graph could be viewed as a form of perceptual *priming*.

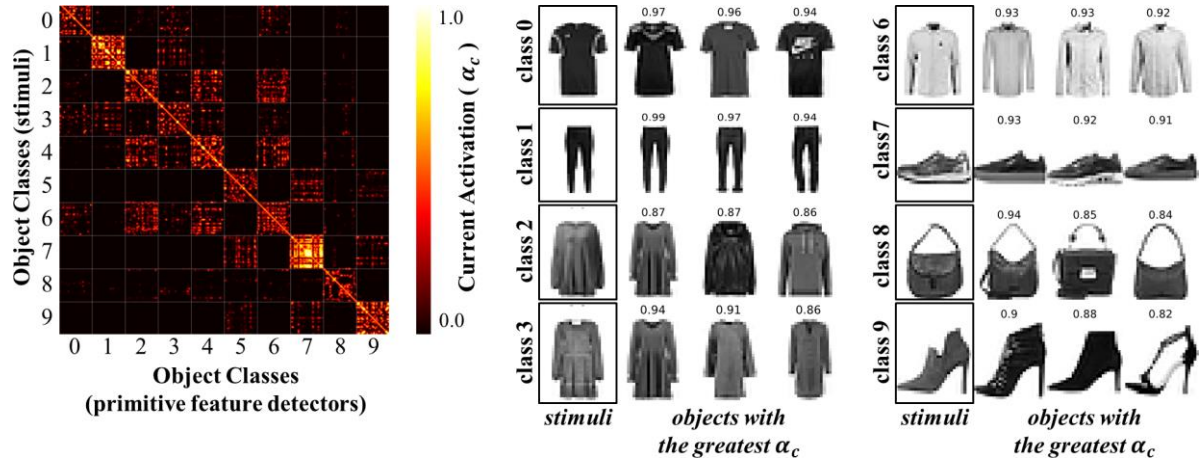


Figure 15. Current activations and resemblance-based best matches. A  $\beta$ -VAE was trained on the Fashion MNIST dataset and used to generate a heatmap of current activations (left) for 250 randomly sampled images. Objects were grouped in the heatmap by their object classes (25 images per class). Gridlines were added to help with visualizing category boundaries. Example sensory stimuli and the objects receiving the most current activation ( $\alpha_c$ ) from those stimuli are also shown (right).

The heatmap (Figure 15, Left Panel) shows that *intra*-category activations were always greater (on average) than *inter*-category activations. Furthermore, categories containing items that resemble one another—for example, pullovers (category 2), coats (category 4), and shirts (category 6)—generated higher inter-category current activations than categories containing dissimilar items. In fact, sensory representations corresponding to dissimilar categories generated very little current activation; for example, the only sensory representations receiving current activations for images of ankle boots were those corresponding to other items of footwear.

Figure 15, Right Panel shows a set of images functioning as incoming sensory stimuli. These sensory stimuli were encoded using the  $\beta$ -BAE’s recognition network and used to activate modal nodes in PAM (i.e., previously learned sensory representations). The set of images associated with the most highly activated modal nodes are shown alongside the sensory stimuli that activated them. Figure 15, Right Panel demonstrates that the most highly activated sensory

representations corresponded to objects that shared similar shapes, textural features, and luminance (i.e., they resemble the sensory stimuli that activated them).

Based on the above observations, I contend that a  $\beta$ -VAE-based implementation of LIDA's Sensory Memory module can learn sensory representations that are analogical. Furthermore, since these sensory representations are used as the basis for all of PAM's grounded concept representations, their analogical property will apply to all grounded representations in PAM.

### ***Grounded***

Sensory representations are non-symbolic representations generated by Sensory Memory's  $\beta$ -VAE recognition networks. They encode the modality-specific, low-level features of environmental stimuli. They are analogical (shown earlier)—capable of serving as proxies for the environmental stimuli they represent. And they can be used to activate other (similar) sensory representations using a resemblance-based comparison (e.g., cosine similarity). Therefore, *sensory representations are grounding representations* (cf. Harnad's iconic representations, Chapter 2).

Elementary grounded concept representations combine one or more *modal* nodes (with associated sensory representations) and a coordinating *amodal* node using referential (activation) links. Referential links support the spread of activation from modal nodes to their connected amodal symbols. This allows the activations associated with those amodal symbols to covary with the patterns of activation of their constituent sensory representations. Assuming Sensory Memory's sensory representations reflect the features of environmental stimuli, and Sensory Memory uses those sensory representations to activate modal nodes in PAM, then elementary

grounded concept representations will also reflect those features in their patterns of activation. As such, elementary grounded concept representations are grounded in their sensory representations; furthermore, *elementary grounded concept representations are grounding representations*. Through induction, one could now show that *all* of LIDA’s grounded concept representations are grounded (via chains of referential links from elementary grounded concept representations).

### ***Generative***

I claim that an implementation has learned generative representations (e.g., sensory representations) if and only if those allow the creation of modal simulations that are *recognizable by the implementation’s system*. For example, given a mentally simulated shirt, shoe, or bag, it should be recognizable as such by that system’s perceptual processes. To demonstrate that this property holds for the  $\beta$ -VAE-based implementation described here, I generated mental simulations for the same 250 randomly selected images that were used to generate the heatmap in Figure 15 (Left Panel). I then used these mental simulations to activate the  $\beta$ -VAEs recognition network *as if they were incoming sensory stimuli*. I generated corresponding sensory representations and calculated the pairwise  $\alpha_c$  as before. The resulting  $\alpha_c$  heatmap (not shown) looked very similar to the heatmap shown in Figure 15 with only slightly more inter-class noise. This demonstrates that the mental simulations produced by the  $\beta$ -VAEs generative networks (and, by extension, simulator structure building codelets) are sufficient to support the correct perception of those “virtual” stimuli.

Analytically, it is self-evident that this property should hold for a  $\beta$ -VAE-based implementation.  $\beta$ -VAE’s latent representations are *designed* to be generative with respect to its

generative network.<sup>24</sup> This is a consequence of its encoder-decoder architecture and its loss function that attempts to minimize reconstruction errors. While the quality of its reconstructions (simulations) will vary (e.g., depending on the  $\beta$ -VAE's layered topology and the  $\beta$  value used), its latent representations will almost certainly be generative after sufficient training.<sup>25</sup> Therefore, LIDA's sensory representations, which are based on these latent representations, will also be generative.

Figure 16 demonstrates this property visually with respect to the  $\beta$ -VAE trained earlier on Fashion MNIST. It shows a set of “real” sensory stimuli and their corresponding mental simulations (i.e., “virtual” sensory stimuli). Notice that these simulations resemble their corresponding “real” stimuli. Also notice that these simulations fail to capture many of the details present in the original images; for example, the buttons on shirts and stripes on shoes are completely missing (i.e., they are “partial” and “indeterminate”; see Barsalou, 1999). Finally, notice that some of the simulated objects have slightly different shapes and orientations than their real counterparts—for example, the shape of handbag and the orientation of the pant legs. Therefore, there is some evidence that this  $\beta$ -VAE has generalized over its inputs rather than simply memorizing those details.

---

<sup>24</sup> This is a further example that mental representations must be understood with respect to the cognitive processes that are intended to interpret them. A latent representation that is “generative” with respect to one  $\beta$ -VAE's generative network will likely produce incomprehensible, incoherent reconstructions using a different  $\beta$ -VAE's generative network.

<sup>25</sup> The only caveat to this is that a bad random initialization of a  $\beta$ -VAE's parameters could cause a gradient descent-based optimizer to become prematurely stuck in a local optimum for which this property fails; therefore, it is always best to experimentally verify these properties.

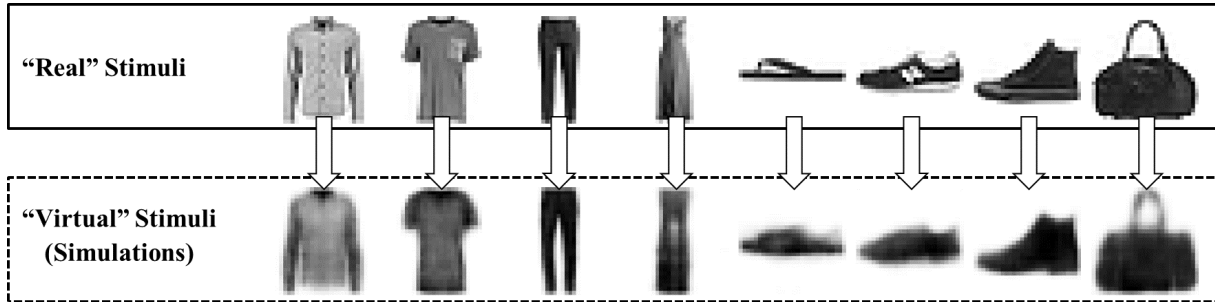


Figure 16. Stimuli and their mental simulations (Fashion MNIST).

### *Perceptual*

Grounded concept representations are perceptual representations—i.e., they are learned into Perceptual Associative Memory in support of recognition. Therefore, there is no disconnect between a LIDA agent’s sensory/perceptual systems and its conceptual representations.

Grounded concept representations can cue grounded concepts in PAM either directly (using sensory representations and structural comparisons) or indirectly (through mediating mental simulations). The latter (simulation-based mechanism) should be preferred in most cases.

### **Discussion**

This chapter detailed implementations for many of ES-Hybrid foundational components in LIDA. These included sensory representations (cf. sensorimotor representations), grounded concept representations, referential and non-referential associations, the generative processes supporting mental simulation (i.e., simulator SBCs), multimodal perception, and “conscious” perceptual/conceptual learning. While there are many ways of computationally realizing this functionality within LIDA, the combination of modality-specific  $\beta$ -VAEs and a content-addressable activation graph provides a powerful and flexibility way of satisfying ES-Hybrid’s representational requirements that should work well for many environments.

Previous LIDA research has attempted to implement grounded cognition in LIDA using a combination of Modular Composite Representation (MCR) vectors (Snaider & Franklin, 2014a, 2014b) and a Hierarchical Temporal Memory (HTM; Hawkins et al., 2010) based Sensory Memory module (see Agrawal et al., 2018). While MCR vectors can be analogical and perceptual, they are not generative; therefore, it is unclear how to use them as the representational basis for modal mental simulations and, by extension, embodied, simulation-based LIDA agents.

Many aspects of the account presented here accord well with Barsalou's theory of Perceptual Symbol Systems (Barsalou, 1999; Barsalou et al., 2003). Sensory Memory's collection of  $\beta$ -VAEs combined with Perceptual Associative Memory's spreading activation graph can be viewed as the "shared associative networks" (Barsalou, 1999, p. 579) that Barsalou suggests as a possible implementation strategy for perceptual symbols. The use of "simulator" structure building codelets to support the generation of mental simulations draws further inspiration from Barsalou. And the cognitive "object" maps described in this chapter are functionally similar to simplified versions of Barsalou's (1999) modal frames. Furthermore, many of the properties Barsalou ascribes to his perceptual symbols (see Chapter 2, Perceptual Symbol Systems) hold for grounded concept representations presented here. Specifically, they are analogical, not complete recordings, partial and indeterminate, capable of designating multiple referents, and dynamic.

However, there are numerous differences as well. Most notably, Barsalou's theory of Perceptual Symbol Systems has been described as an "eliminativist" position (see Goldstone & Barsalou, 1998) that argues that *modal* representations and the cognitive processes that operate



on them are sufficient to account for all of cognition. By contrast, the conceptual and computational implementations described here make extensive use of *amodal* representations. I contend that these amodal representations serve many critical functions that are difficult (and perhaps impossible) to accommodate with purely modal representations. Replacing amodal representations with modal representations, such as cross-modal conjunctive representations (CCRs; see Binder, 2016), which store the “statistically likely features” extracted from category exemplars or multimodal compressed representations (Barsalou, 2016a, p. 1133), would undermine many of those benefits.

A second difference between the current account and Perceptual Symbol Systems is that Barsalou’s (1999) “simulators” are defined as the *combination* of knowledge and generative processes needed to represent a *single* concept. The current implementation separates these concerns (see the earlier section on Simulator Structure Building Codelets). I do this for several reasons, chief among these is computational and conceptual simplicity. Coordinating the activity of the myriad of largely independent, generative processes needed to support a rich conceptual system would be extraordinarily challenging in software. Moreover, a one-to-many relationship between generative processes and the concepts they simulate is more consistent with Kosslyn et al.’s (1988) experimental observations that simulation appears to use sequential generation processes. If the generation of each constituent part of a multi-part object were handled by separate processes operating in parallel, then additional synchronization would be needed to produce those observed experimental results. Therefore, the implementation described here provides a more parsimonious explanation for those experimental results (*ceteris paribus*).

## Chapter 6

### Action-Based Mental Simulation and Motor Cognition

*The overriding task of Mind is to produce the next action....* The various cognitive functions—recognizing, categorizing, recalling, inferencing, planning—all ultimately serve what to do next. (Franklin, 1995, p. 412)

Chapter 5 focused on perception, the representation of grounded concepts, and the generative processes that support mental simulation. This chapter builds on that earlier groundwork, adding aspects of *motor cognition* (see Chapter 2) to LIDA. Specifically, this chapter focuses on developing *action-based* mental simulations and other action-oriented aspects of LIDA’s cognitive cycle.

Procedural Memory and Action Selection are the modules that most directly answer the question “What do I do next?” (Franklin et al., 2016, p. 106) in LIDA. Procedural Memory contains an *internal model* of the predicted consequences of an agent’s actions. And Action Selection uses situationally relevant portions of that model to select an agent’s next action.

Historically, the knowledge in LIDA’s Procedural Memory module was considered to be “never conscious” (see Franklin & Baars, 2010) knowledge. It supported the selection of actions, but its knowledge representations (i.e., schemes) were never consciously accessible to an agent—that is, they were inaccessible from LIDA’s Current Situational Model (CSM).

Building on Jeannerod’s theory of motor cognition (Jeannerod, 2001, 2006), I suggest how “covert” actions (see Chapter 2, Motor Cognition) could support *action-based* mental simulations. Action-based mental simulations are generated from the predicted consequences of

*internally* executed behaviors. Specifically, these mental simulations update LIDA's CSM to reflect the expected outcomes of an agent's selected behaviors. This capacity to internally executed behaviors and generate mental simulations from them is learned by internalizing the consequences of *externally* executed behaviors.

Action-based mental simulations are the basis for volitional/intentional mental imagery (see Chapter 7), and they are likely pervasive in cognitive activities such as deliberation and planning. As such, conscious reasoning (e.g., volitional and consciously mediated offline cognition) could be viewed as a skill that is developed over time through relevant environmental interactions (Bartlett, 1958).

In support of action-based mental simulations and the internal execution of selected behaviors, this chapter details a new implementation for LIDA's Procedural Memory and Action Selection modules. In addition to laying the foundation for motor cognition in LIDA, it advances LIDA's procedural learning, behavior streams, and exploratory action selection. It also suggests a possible implementation for *automatized* action selection (see Chapter 4) based the "overlearned" components of reliable behavior streams.

My implementation is based on an enhanced LIDA-compatible extension of Drescher's (1991) *schema mechanism*. Where Drescher was primarily focused on explicating the very early, sensorimotor stages of Piaget's constructivist theory of childhood development (Piaget, 1952, 1954), I expand that scope to include the mental simulation of actions and their environmental consequences. The capacity to perform action-based mental simulations is believed to emerge later in childhood development (e.g., see Molina et al., 2008; Spruijt et al., 2015).

## **Background: The Schema Mechanism**

Drescher (1991) described his *schema mechanism* as a symbolic, “general learning and concept-building mechanism” inspired by Piaget’s theory of *constructivism* (Piaget, 1952, 1954).<sup>1</sup>

According to this theory, human infants initially conceive of their world *exclusively* in terms of sensorimotor activity. However, through continued interactions with the world, children learn that some of their actions affect their sensations, and gradually they can *construct* more sophisticated representations of reality that are grounded in these sensorimotor primitives.

Piaget described this early learning as progressing through a series of developmental stages. Drescher’s schema mechanism is primarily focused on implementing portions of Piaget’s “sensorimotor stage,” which corresponds to the first two years of an infant’s life. An important aspect of this theory, and Drescher’s schema mechanism, is that knowledge builds on previously learned knowledge. In the subsections that follow, I will describe the schema mechanism’s knowledge structures (i.e., mental representations), action selection, and procedural learning algorithm.

### ***Knowledge Representations***

The schema mechanism’s primary mental representation is the *schema*. It is a three-part data structure composed of a *context*, an *action*, and a *result*. Taken together, these components predict what *might* occur if an agent were to execute an action in a given environmental state.

---

<sup>1</sup> While I am primarily interested in the schema mechanism’s capability to learn procedural knowledge, it should be noted that Drescher intended for the schema mechanism to be a more-or-less complete cognitive architecture, featuring procedural and declarative knowledge, as well as action selection.

Contexts and results make assertions about the world (i.e., an agent's environment)—specifically, the state of the world before and after an action is taken. They are not complete descriptions of an environment. Rather, they only specify those state elements that are statistically correlated with an action and its reliable execution. A schema's context and result are encoded using a set of *zero* or more *item assertions*.

*Items* are binary state elements that symbolize aspects of an agent's environment. They can be thought of as propositions or binary feature detectors that correspond to specific environmental conditions. An item's state (On or Off) indicates whether that condition is satisfied in an agent's environment. And an item assertion stipulates a specific value for an item's state.

A schema's context and result contain a *set of item assertions* that describe relevant aspects of an agent's environment prior to and following the execution of the schema's action. For example, given three items  $p$ ,  $q$ , and  $r$ , a schema's context might be encoded as " $\sim pq$ ". This indicates that item  $p$  should be Off,  $q$  should be On, and it makes no claims about the state of item  $r$ ; that is,  $r$  can be On or Off, and this context would still be satisfied. A schema's context is said to be *satisfied* when all of its positive item assertions correspond to items that are On in an agent's environment, and all of its negative (or negated) item assertions correspond to items that are Off in an agent's environment.

The schema mechanism defines two types of items: primitive items and synthetic items. Primitive items are *built-in* symbolic representations that are toggled On or Off by innate cognitive processes (based on an agent's current environmental state). In simple cases, primitive items could be wired directly to an agent's sensors (e.g., collision detectors) to set their values

On or Off. More generally, primitive items function as Boolean functions (predicates) that characterize the observable properties of an agent's environmental state. In contrast, synthetic items are *learned* additions to the schema mechanism's conceptual repertoire.

The schema mechanism creates a synthetic item whenever a schema is found to be unreliable but "locally consistent" (Drescher, 1991, p. 82). Drescher referred to such a schema as the synthetic item's *host schema* and the resulting synthetic item as the host schema's *reifier*. Conceptually, a synthetic item represents some unknown, previously unconceived-of, environmental condition that influences the reliability of its host schema.<sup>2</sup>

To make these ideas more concrete, consider the schema depicted in Figure 17. It specifies an action of moving one's hand to the left (**MOVE-HAND-LEFT**) and a result of one's hand touching something to the left (**HAND-TOUCH-LEFT**). This schema is likely *unreliable* because it fails to specify any conditions that constrain its applicability. Its context is empty (i.e., it asserts nothing), so it is assumed to apply in all situations. However, if there are no objects to an agent's left, that agent should not expect to touch something if its hand were moved to its left. On the other hand, if on some occasion the agent did touch something after moving its hand to the left, then that result is more likely to occur again when repeating the same action (at least over a short duration of time).<sup>3</sup> Such a schema is said to be *locally consistent*. And when the schema

---

<sup>2</sup> Synthetic items are examples of "initially ungrounded" symbols that signify "unknown referents" (see Chapter 3). They are hypothesized environmental conditions that are *inferred* from other environmental regularities. The schema mechanism works backwards from a previously conceived manifestation (e.g., a tactile sensation) and postulates a previously *unconceived-of thing* (e.g., a physical object). This new synthetic item (i.e., concept) is allocated by the schema mechanism and grounded through "the reification of counterfactual assertions" (Drescher, 1991, p. 90).

<sup>3</sup> The assumption being that objects generally remain in approximately the same place over short periods of time.

mechanism detects this local consistency, it creates a new synthetic item (labelled **PALPABLE-OBJECT-LEFT** in Figure 17) with this unreliable schema as its host schema.

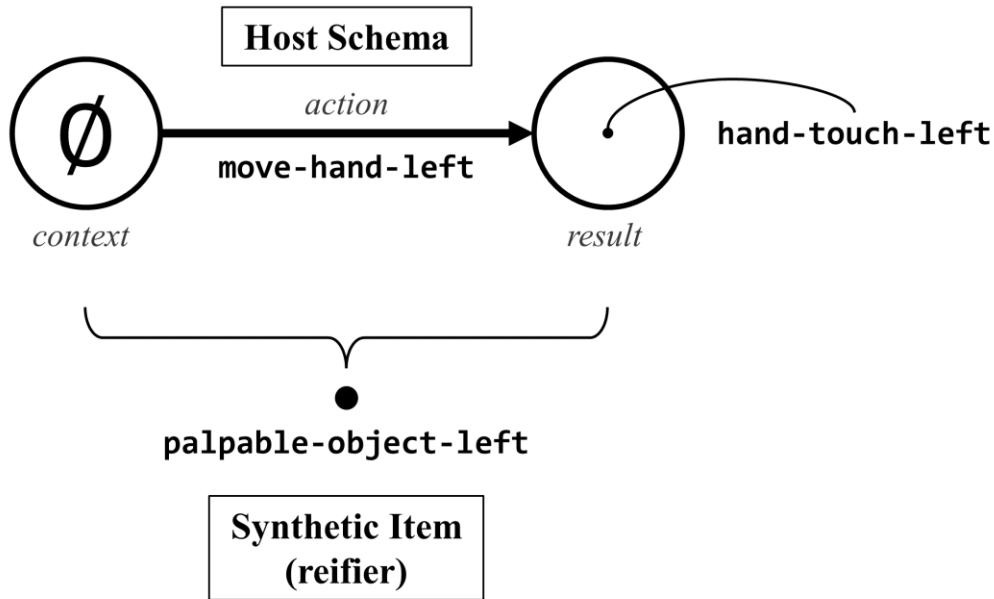


Figure 17. A synthetic item and its host schema. A schema with an empty context (top) specifies that if an agent’s hand were to move to the left (**MOVE-HAND-LEFT**) it would result in the agent’s hand touching an object to its left (**HAND-TOUCH-LEFT**). This schema was identified as an unreliable, but locally consistent schema, resulting in the creation of a synthetic item (bottom). This synthetic item can be interpreted as corresponding to the condition of a palpable object being to the agent’s left (**PALPABLE-OBJECT-LEFT**).

Unlike primitive items—which are considered On or Off based on the state of a sensor or a perceptual process—the schema mechanism determines a synthetic item’s On/Off status based on a set of learned “verification conditions” (see Drescher, 1991, sec. 4.2.2). The simplest of these is based on whether a synthetic item’s host schema’s result occurs after its host schema’s action is taken. If the host schema’s result does occur when its action is taken, then the synthetic item is turned On; if the host schema’s result fails to occur after its action is taken, then the synthetic item is turned Off.

The schema mechanism supports two types of actions: primitive actions and composite actions. *Primitive actions* are built-in actions that are typically “hard-wired” to actuators or controllers that execute those actions. *Composite actions* are learned actions that function like subroutines; specifically, each composite action contains a *controller* that learns *chains of (component) schemas* (see Figure 18) that lead to a specific *goal state*.

Once learned, composite actions provide an abstraction over the details of how these goal states are reached, allowing the schema mechanism to learn further results that follow from achieving those goal states. For example, consider a composite action with a goal state of **ON(SWITCH)**, that is, “a switch being in its on position.” This might result in “the lights being on”—**ON(LIGHTS)**—in the context of a kitchen, that is, **IN(KITCHEN)**; however, but it might result in “a fan being on”—**ON(FAN)**— in the context of a bedroom, that is, **IN(BEDROOM)**. The schema mechanism could learn these regularities as two different schemas with the same composite action: **IN(KITCHEN)/ON(SWITCH)/ON(LIGHTS)** and **IN(BEDROOM)/ON(SWITCH)/ON(FAN)**.<sup>4</sup>

---

<sup>4</sup> Drescher (1991) often depicted schemas as “CONTEXT/ACTION/RESULT” with forward slashes separating each component. I follow that convention here.



## Chain of Schemas

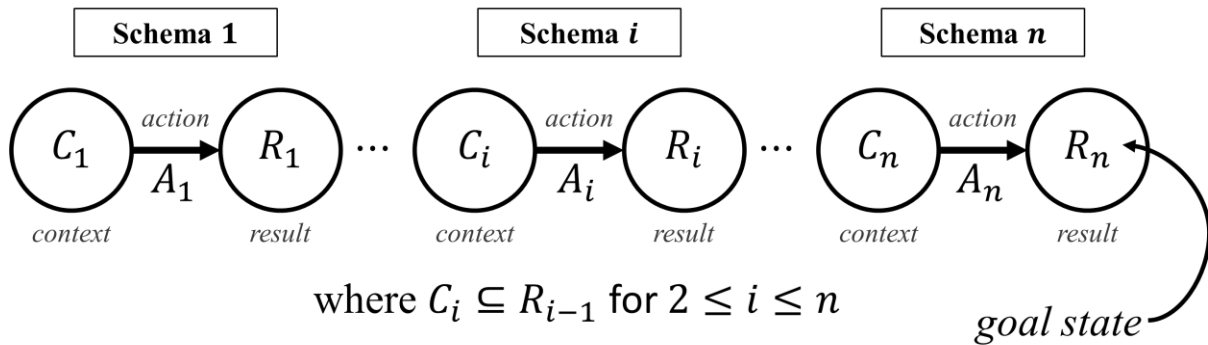


Figure 18. An illustration of a single chain of schemas. One or more schemas can chain together (e.g., within a composite action's controller) when one schema's context is a proper subset of another schema's result, forming a link in the chain. For composite actions, the terminal link in a schema chain will have the composite action's goal state as its result. Note that composite actions typically contain *multiple* chains of schemas that terminate in a single goal state.

Schemas have several associated properties. A schema's *reliability* tracks the probability with which a schema's result occurs when its action is taken. In general, schema's may be arbitrarily unreliable; however, even unreliable schemas are useful because they serve as points of departure for learning more reliable schemas. (Exactly how the schema mechanism achieves this will be discussed later in this chapter.) Other parameters associated with each schema include its (average) *duration*, (average) *cost*, and *overriding conditions* (see Drescher, 1991, p. 55).

### **Action Selection**

The schema mechanism selects a single *applicable* schema at each discrete time step. A schema is said to be applicable when its context is satisfied and none of its overriding conditions occur. Schema selection is based on a multi-faceted appraisal of each applicable schema. The criteria used to evaluate each schema can be broadly characterized as being in service of *goal-pursuit* or *exploration* (cf. exploration and exploitation in reinforcement learning; see Sutton & Barto,

2018). The schema mechanism adjusts the relative importance of these criteria over time in order to maintain a “cyclic balance” between goal-pursuit and exploration (see Drescher, 1991, p. 66).

A schema’s goal-pursuit importance is primarily determined by the primitive, delegated, and instrumental values of the items in its result. *Primitive value* is a built-in<sup>5</sup> measure of the (positive or negative) desirability of an item. *Delegated value* is an acquired measure of the (positive or negative) desirability of an item. And *instrumental value* is a transient, goal-specific measure of the current utility of an item.<sup>6</sup>

The primary criteria influencing a schema’s exploratory importance are what Drescher called hysteresis and habituation (see Drescher, 1991, p. 66). *Hysteresis* is the tendency to re-select recently selected schemas, which provides a temporary focus of attention around a small set of schemas. *Habituation* is a temporary devaluation (partial suppression) of schemas that have been recently selected many times. In addition to hysteresis and habituation, the schema mechanism also includes other exploratory mechanisms that encourage the selection of schemas with underrepresented actions.

The selection of schemas with composite actions introduces additional complications. While the initial selection of a composite-action schema proceeds exactly like a schema with a primitive action, they are treated differently after their selection. For example, immediately after the selection of a schema with a composite action, the schema mechanism will choose one of its

---

<sup>5</sup> Since primitive values are “built-in” values, they are only associated with primitive items (not synthetic items, which are learned).

<sup>6</sup> An item is said to have *instrumental value*—with respect to a specific chain of schemas leading to a goal state—if turning that item On would help satisfy the context of the next schema in that schema chain.

composite action's applicable *component schemas* for execution. Applicable component schemas are selected based on their *proximity* to the composite action's goal state, rather than their goal-pursuit or exploratory value. If the selected component schema is itself a composite-action schema, then the schema mechanism will continue to recurse in this way until a schema with a primitive action is found.

On subsequent selection events, a previously selected composite action schema—referred to as the *pending schema*—must still compete for re-selection like any other schema; however, it is given additional importance that encourages its re-selection. Drescher stated,

The mechanism grants a pending schema enhanced importance for selection, so that the schema will likely be re-selected until its completion, unless some far more important opportunity arises. Hence, there is a kind of focus of attention that deters wild thrashing from one never-completed action to another, while still allowing interruption for a good reason. (Drescher, 1991, p. 62)

The current pending schema is aborted if it fails to make progress towards its goal state after a maximum expected execution time, or, if at any time following its initial selection, the pending schema has no applicable component schemas.

### ***Procedural Learning***

The schema mechanism learns new schemas by *marginal attribution*—an empirical learning algorithm that is designed to incrementally learn chains of reliable schemas. A fundamental feature of marginal attribution is that new knowledge builds on prior knowledge. That is, new schemas are incrementally constructed from existing schemas through a process called *spin-off*.

The schema mechanism is initialized with a set of primitive actions and primitive items. For each of its primitive actions, the schema mechanism constructs a *bare schema*—an action-only schema that serves as a “point of departure” for the discovery of action consequences. From this set of bare schemas, the schema mechanism’s marginal attribution process attempts to identify the statistically significant effects of an agent’s actions. If an item is significantly more likely to turn On when an agent executes an action, then a *result spin-off* schema is created containing that *relevant item* in its result. This spin-off schema’s action will be the same as its originating bare schema’s action, and its context will be empty. Similarly, if an item is significantly more likely to turn Off when an agent executes an action, then a result spin-off schema is created containing the *negation* of that relevant item.

Once the schema mechanism learns one or more result spin-offs, it then attempts to identify context conditions that will make these result spin-offs more *reliable*. If a schema’s result is significantly more likely to occur when a particular item is On prior to its action execution, then a *context spin-off* is created containing that relevant item in its context. This new spin-off schema’s action *and* result will be the same as its originating schema’s action and result. Similarly, if a schema’s result is significantly more likely to occur when a particular item is Off prior to its action’s execution, then the schema mechanism will create a context spin-off schema with a negated assertion for that item.

For each context spin-off created in this way, additional context spin-offs can be created from those schemas—*new knowledge builds on previous knowledge*. This supports the construction of schemas with composite contexts that function like conjunctions (logical ANDs)

of items.<sup>7</sup> Once a conjunction of relevant items has been identified (via a context spin-off), it is eligible for inclusion in a result spin-off (but not before)<sup>8</sup>.

Schemas require two additional “extended” components to support the identification of relevant items: an *extended context* and an *extended result*. These components tabulate *item-level* correlation statistics with respect to each schema. Extended context statistics, which support context spin-offs, are eligible to be updated whenever a schema is explicitly or implicitly *activated*. A schema is *explicitly activated* when it is selected for execution. A schema is *implicitly activated* whenever it shares the same action as the explicitly activated schema, and its context is satisfied.

The schema mechanism updates the *reliability* associated with *all* implicitly or explicitly activated schemas based on the environmental state that follows action execution. If the resulting state satisfies an activated schema’s result, then its reliability is increased; otherwise, its reliability is decreased.

The schema mechanism maintains numerous other (learned) parameters that are periodically updated including a schema’s *cost* and *duration*; an item’s *generality*, *accessibility*,

---

<sup>7</sup> The schema mechanism does not *explicitly* support disjunctions of items (logical ORs over sets of items); however, when multiple schemas contain the same context and action, but different results, these schemas *implicitly* function as a disjunction over their results. Similar, when two or more schemas have the same action and result, but different contexts, these sets of schemas function as implicit disjunctions with respect to their contexts.

<sup>8</sup> This is a subtle, but important point. The schema mechanism does not support the incremental learning of composite results, like it does for composite contexts. Drescher explained, “to prevent the explosive proliferation of such combinations,... only a schema with an empty result [i.e., a bare schema] can spin off a schema with a new result item” (1991, pp. 78–79). While composite results are possible, they are limited to those conjunctions of conditions that appear in known composite contexts. That is, the schema mechanism focuses on composite results that, when obtained, will satisfy the context of a known context-spinoff. This encourages the learning of schema chains.

and *delegated value*; and a component schema's *proximity* with respect to its composite action's goal state. These details have been omitted for brevity. I have also purposefully glossed over the specific processes that govern the learning of composite actions, synthetic items, and a schema's overriding conditions.

## Implementation

This section details an enhanced version of Drescher's (1991) schema mechanism, which I use to implement LIDA's Procedural Memory and Action Selection modules. In addition to supporting simulation-based cognition, this implementation solves several of LIDA's open research problems. It implements *instructionist* procedural learning—that is, *when* and *how* new procedural knowledge is acquired.<sup>9</sup> It implements *behavior streams* (i.e., “action plans”; see Ramamurthy et al., 2001), which have long existed in LIDA's conceptual model, but lacked a concrete implementation. And it implements exploratory action selection<sup>10</sup>, which has yet to be considered in the LIDA literature.

While Drescher's (1991) schema mechanism was intended to be a *complete* cognitive system—supporting notions of procedural memory, declarative memory, action, and perception—the version proposed here only implements a small portion of LIDA; specifically, its Procedural Memory and Action Selection modules. As a result, some of the representations used by this implementation are managed by LIDA modules and processes that are external to it

---

<sup>9</sup> Recall that *instructionist learning* involves learning new representations, and *selectionist learning* involves reinforcing existing representations (e.g., their associated parameters; see Edelman, 1987). Instructionist procedural learning has never been computationally implemented in LIDA (e.g., in LIDA's software framework; see Snaider et al., 2011), and its conceptual implementation is still underdeveloped.

<sup>10</sup> The exploration vs. exploitation problem, which is often discussed in the reinforcement learning literature (see Sutton & Barto, 2018) but has never been addressed in the LIDA literature.

and outside of its control. In other cases, the schema mechanism's representations (e.g., synthetic items) and processes (e.g., conceptual learning) needed to be replaced by LIDA counterparts that are often quite different (both conceptually and functionally).

As such, I begin this section by detailing how LIDA's representations and processes can be reconciled with the schema mechanism. While this is largely a mapping exercise, some deep analyses and occasional compromises were needed. Table 1 lists the terminology that is roughly analogous between the two cognitive models. (It also indicates terms that have no clear correspondence.)

Next, I detail my schema-mechanism-based implementation of LIDA's Action Selection and Procedural Memory modules. This implementation is enhanced to support simulation-based cognition; specifically, the internal (covert) execution of behaviors and action-based mental simulations (i.e., motor cognition; see Chapter 2). These new capabilities are used in Chapter 7 to implement volitional/intentional mental imagery in LIDA.

Table 1. Terminological comparison of Drescher’s Schema Mechanism and LIDA. Terms that appear in the same row are roughly analogous, while asterisks (\*) indicate terms without corresponding concepts.

LIDA	SCHEMA MECHANISM
schemes	schemas
base-level activation (of schemes)	reliability (of schemas)
*	extended context / extended result (of schemas)
amodal nodes	items
amodal nodes (learned)	synthetic items
node structures	sets of items
affective valence	primitive value
incentive salience	delegated value
*	instrumental value
behavior streams	composite actions
sensory representations	*
modal node / feeling node	*
templated schemes	*
selectable behaviors	applicable schemas
internal actions	*

### ***Reconciling LIDA with the Schema Mechanism***

**Schemes and Schemas.** *Schemes* are the primary data structures used in LIDA’s Procedural Memory module (see Chapter 4). Conceptually, they are similar to Drescher’s schemas since they were inspired by Drescher’s schemas<sup>11</sup>. Both are composed of a *context*, an *action*, and a *result*, and the meaning of these components is essentially the same. However, despite these surface similarities, their details often differ.

Traditionally, LIDA’s schemes encode their contexts and results as *node structures* (see Chapter 4). Node structures are directed graphs containing one or more nodes, and zero or more

---

<sup>11</sup> The decision was made to name LIDA’s units of procedural knowledge “schemes” rather than “schemas” to avoid misleading readers into thinking that LIDA’s version of these data structures represented forms of non-procedural knowledge (such as forms of declarative and conceptual knowledge) that applied to Piaget’s more expansive formulation (S. Franklin, personal communication, July 29<sup>th</sup>, 2022).



directed links. Nodes represent concepts or instances of concepts, and links represent the relationships between them. Node structures are *structured* composite representations—i.e., their links carry additional semantic information that can influence their interpretation.

In contrast, Drescher's schemas represent their contexts and results as *sets of item assertions*. Item assertions indicate the presence or absence of environmental conditions, and each item assertion is logically independent of one another. Consequently, the contexts and results encoded in Drescher's schemas are *unstructured* composite representations.

Unfortunately, it is impossible to map structured representations to unstructured representations without losing information. Therefore, a pragmatic compromise is needed to reconcile LIDA's node structures with the schema mechanism's knowledge representations and learning processes.

Individually, LIDA's *amodal nodes* (see Chapter 5) are similar to the schema mechanism's items. And their inclusion in, or omission from, a conscious broadcast can be used to determine an item's current state (On or Off). This suggests that Procedural Memory's contexts and results could be encoded as *sets of node assertions* that are activated based on the presence or absence of amodal nodes in a conscious broadcast. This, in turn, could support the instantiation of situationally relevant schemes.<sup>12</sup>

---

<sup>12</sup> While this differs from LIDA's conceptual model (Franklin et al., 2016, sec. 5.6)—which assumes both the ability to structurally compare node structures and the ability to perform structure-based, instructionist procedural learning (i.e., adding or deleting structures)—similar simplifications have been employed in all of LIDA's computational implementations. For example, LIDA's Java Framework uses a node-based comparison to activate schemes (links are ignored).

**Base-Level Activation and Reliability.** LIDA’s schemes have *base-level activations*. Franklin et al. (2016, p. 119) described a scheme’s base-level activation as a measure of the likelihood that the execution of a scheme’s action will achieve its result if its context is satisfied. Based on this description, base-level activation is analogous to a schema’s *reliability*. Furthermore, the selectionist procedural learning algorithm that LIDA uses to update its schemes’ base-level activations is similar to the schema mechanism’s reliability update rule.<sup>13</sup>

That said, base-level activations (LIDA) and reliabilities (schema mechanism) have subtly different mechanics due to LIDA’s comprehensive decay processes (see Franklin et al., 2016, sec. 4.6). Since a scheme’s base-level activation decays (albeit slowly), it incorporates notions of frequency and recency that are not present in the classical schema mechanism’s reliabilities. As such, an otherwise reliable scheme (in LIDA) may be interpreted as being unreliable if its action is rarely executed, or it has been a while since its last execution.

These differences do not negatively impact the schema mechanism’s operations. On the contrary, the additional functionality afforded by base-level activations can be seen as an enhancement over the original schema mechanism’s reliabilities. Therefore, base-level activation can be used as a “drop-in replacement” for the schema mechanism’s notion of reliability.

**Extended Contexts and Extended Results.** Drescher’s schemas include *extended contexts* and *extended results*—data structures that tabulate item-level statistics in support of marginal attribution. They have no counterparts in LIDA. My implementation of Procedural Memory—in

---

<sup>13</sup> With important differences that will be discussed later in this chapter (see Procedural Memory, Selectionist Procedural Learning).

particular, instructionist procedural learning—requires that they be added as components of LIDA’s schemes.

**Amodal Nodes and Items.** I previously mentioned that LIDA’s amodal nodes are similar to the schema mechanism’s items. (That statement is only partially true.) Representationally, amodal nodes and items serve the same function: they symbolize “things” in an agent’s environment (objects, entities, situations, events, etc.). However, amodal nodes (LIDA) and items (schema mechanism) are learned and processed in very different ways.

Recall that the schema mechanism has two types of items: primitive and synthetic. Primitive items are *never* learned. They are innate (built-in) concepts that are hard-wired to feature detectors and perceptual processes. Synthetic items, on the other hand, are *learned* concepts that rely on their host schemas for activation. They are detached from an agent’s sensory and perceptual processes (e.g., feature detectors and other concepts), and their meaning depends entirely on their host schemas. Thus, synthetic items are largely incompatible with LIDA’s current notions of (grounded) concept representation (see Chapter 5).

Reconciling the schema mechanism’s purely symbolic conceptual system with LIDA’s more capable, hybrid conceptual system (see Chapter 5) requires a fundamental change to the schema mechanism’s concept representations and representational learning processes.

Specifically,

- (1) *new* primitive items must be supported,
- (2) synthetic items must be removed, and
- (3) concept learning must be externalized from the schema mechanism.

Supporting new primitive items would have been an impossibility for the classical schema mechanism. It lacked a way to experientially connect new items to an agent's sensory and perceptual systems; therefore, it had no means of updating their internal states—i.e., turning them On or Off—based on an agent's current environment. Fortunately, LIDA has solved this problem (see Chapter 5).

Structure building codelets create new amodal nodes based on incoming sensory stimuli. These nodes are (referentially or non-referentially; see Chapter 3) connected to modal nodes in LIDA's Perceptual Associative Memory module. Upon receiving a conscious broadcast containing never-before-seen amodal nodes, LIDA's Procedural Memory module will expand its extended contexts and extended results to include a new slot for each such amodal node. Procedural Memory will then begin tabulating correlation statistics for these new nodes. Procedural memory also maintains a reference to each amodal node, which will be used to support instructionist procedural learning—i.e., context and result *spin-offs*.

**Sensory Representations, Modal Nodes, and Feeling Nodes.** LIDA's sensory representations, modal nodes, and feeling nodes function as *non-symbolic* representations. They support the grounding and activation of *amodal* nodes, and the learning of motivation-related parameters (i.e., incentive saliences) associated with those amodal nodes. The schema mechanism was not designed to work directly with non-symbolic representations, and, in general, sensory stimuli, sensory representations (e.g., modal probability distributions; see Chapter 5), and the modal nodes that encapsulate them should be *excluded* from LIDA's schemes.

For similar reasons, feeling nodes (see Chapter 4) should be excluded from Procedural Memory's schemes. Feeling nodes are not concept representations.<sup>14</sup> They do not serve to identify environmental stimuli. They quantify an agent's immediate hedonic responses—feelings of liking or disliking—elicited by those stimuli.<sup>15</sup> Feeling nodes function like modal nodes, and, in many cases, they could be implemented as such. Separate perceptual/conceptual (amodal) nodes must be instantiated for an agent to recognize the external stimuli that provoked those feelings (e.g., sweet foods). And it is those conceptual representations that should be included in schemes (not feeling nodes).

In summary: LIDA's schemes' contexts and results should be composed exclusively from *amodal* node assertions. Modal nodes, feeling nodes, sensory representations, and sensory stimuli should be limited to the data structures and processes that support other long- and short-term memory modules (e.g., Perceptual Associative Memory and the Current Situational Model).

**Affective Valence and Primitive Value.** LIDA's motivational system is grounded in *affective valences*—parameters associated with LIDA's feeling nodes (see Chapter 4). A feeling node's total activation determines the magnitude of its affective valence, and its valence sign (positive or negative) determines whether that sensation is interpreted as pleasant or unpleasant. Affective valence represents an agent's hedonic response to environmental stimuli—for example, the

---

<sup>14</sup> This is an important point. While feeling nodes are differentially activated in response to particular stimuli, such as sweets, they do not serve as percepts that identify the sweetness of substances. Instead, feeling nodes quantify an agent's *liking* or *disliking* of that stimulus—that is, how the agent “feels” in response to sweet foods. A separate (perceptual) node must be instantiated for the agent to recognize that an external stimulus contained a “sweet substance.” It is this percept that should be included in schemes (not feeling nodes).

<sup>15</sup> The same experience can elicit different hedonic responses (e.g., the magnitude of those feelings may change), depending on the dynamics in Perceptual Associative Memory (PAM) and the situational contexts surrounding that event.

“liking” or “disliking” of bodily sensations (e.g., hunger and satiety) and events (e.g., eating sweet or bitter foods). However, feeling nodes do not *identify* the environmental stimuli that evoked those affective responses. Affective valence is merely elicited by them. Consequently, feeling nodes are typically *co-activated* with the grounded concept representations (see Chapter 5) that identify those environmental stimuli. (This fact enables associative learning between concept representations and their co-occurring feelings in a conscious broadcast.)

By comparison, the schema mechanism’s motivational system is grounded in *primitive values*—parameters associated with the schema mechanism’s primitive items (see Background: The Schema Mechanism). Primitive values represent the intrinsic desirability agents associate with their primitive items’ environmental referents. Unlike LIDA’s affective valences (which are attached to feeling nodes), primitive values are attached to representations that symbolize those environmental referents (rather than an agent’s affective responses to them).

Consequently, primitive values can be included as components of contexts and results, and they can *directly* support goal-directed action selection. This differs from LIDA’s feelings nodes, which are excluded from contexts and results, and, as a result, can only *indirectly* support goal-directed action selection (via their role in incentive salience learning).

These differences between affective valence (LIDA) and primitive values (schema mechanism) are largely irrelevant in the context of motivational learning. Both affective valence and primitive value function as built-in “reward signals” (cf. rewards in reinforcement learning; Sutton & Barto, 2018) that enable concept representations to acquire their derived motivational values—incentive saliences (LIDA) and delegated values (schema mechanism). That is, with

respect to motivational learning, LIDA's affective valences are analogous to the schema mechanism's primitive values, and they can replace them as motivational primitives.

**Incentive Saliency and Delegated Value.** Both LIDA's and the schema mechanism's conceptual representations support *learned* motivational values. These are called *incentive saliency* and *delegated value*, respectively.

Incentive saliency (LIDA) quantifies the current value an agent places on an event occurring in the environment. It functions as a measure of the “wanting” or “dreading” associated with that event. Incentive saliency combines an event's historical desirability (base-level incentive saliency) with the realities of an agent's current situation (current incentive saliency). For example, while the thought of drinking water may generally have low-to-moderate desirability (base-level incentive saliency), intense feelings of thirst can greatly increase water's desirability (current incentive saliency).

By comparison, Drescher (1991) described an item's delegated value as accruing to states that generally facilitate the acquisition of other things of value. That is, items acquire delegated value from the things they help to achieve rather than from any intrinsic (primitive) value that those items may possess. This measure of desirability (or utility) is not goal-specific; it is a general facet of that item with respect to all of the goals it facilitates.<sup>16</sup> In a similar way, one might say that LIDA's *base-level* incentive saliency accrues to nodes that frequently lead to

---

<sup>16</sup> Drescher (1991) gave an example involving a young child and its parent to illustrate delegated value. The parent acquires delegated value for that child because having its parent in close proximity has general utility. Regardless of the child's specific needs, its parent is there to facilitate their satisfaction.

events an agent “likes.”<sup>17</sup> As such, base-level incentive salience and delegated value are analogous (though slightly different) concepts. The primary functional difference being that incentive salience is a *context-sensitive* measure of desirability, whereas delegated value is context-agnostic.

While incentive salience (LIDA) and delegated value (schema mechanism) as subtly different conceptually and functionally, they are effectively identical with respect to their support of goal-directed action selection. Moreover, they are both derived from more basic motivational constructs (i.e., affective valence and primitive value). In short, without delving into other irrelevant computational differences here, LIDA’s incentive saliences can replace the schema mechanism’s delegated values without introducing computational or conceptual difficulties.

A general enhancement (related to motivational learning) that I have included in my computational implementation is the use of temporal-difference (TD) learning with “replacing” eligibility traces (Singh & Sutton, 1996) to update base-level incentive salience.<sup>18</sup> Specifically, whenever one or more feeling nodes are globally broadcast, Perceptually Associative Memory (PAM) updates the base-level incentive salience associated with *all* amodal nodes that contributed to that “feeling event” (McCall et al., 2020, sec. 5.2). The magnitude and direction of their update is based on (1) the combined affective valence over those broadcast feeling nodes and (2) how *recently* each amodal node was included in a conscious broadcast. This requires an

---

<sup>17</sup> This characterization is based, in part, on the fact that temporal-difference (TD) learning (Sutton & Barto, 2018, Chapter 6) is used to update the base-level incentive saliences associated with LIDA’s nodes (see McCall et al., 2020, sec. 5.3). As such, the base-level incentive saliences associated with nodes for objects, entities, situations, events that contribute to acquiring those things an agent “likes” will also acquire some base-level incentive salience.

<sup>18</sup> McCall et al. listed *eligibility traces* as an “avenue of future research” (McCall et al., 2020, p. 62) with respect to LIDA’s motivational learning. This avenue has been explored as part of this work.



additional parameter (i.e., an eligibility trace) be added to each PAM node to track the recency of its broadcast. It also requires that PAM decays these values slightly following each global broadcast.

**Behavior Streams and Composite Actions.** Ramamurthy et al. (2001, p. 7) defined a *behavior stream* (in LIDA) as a “partially ordered plan which guides [the] execution of behaviors... so as to effect the required transition from the initial state to the goal state.” Compare this with Drescher’s definition of a composite action as an action that is “defined with respect to some goal state; it is the action of bringing about that state.... The means are given by chains of schemas that lead to the goal state from various other states” (1991, p. 59). Based on these descriptions, behavior streams are functionally analogous to composite actions.

Given that behavior streams have never been fully implemented in LIDA, an added benefit of this schema-mechanism-based implementation is that behavior streams would have a clear conceptual and computational implementation (both in terms of its representation, procedural learning, and action selection). This implementation may also pave the way for an implementation of LIDA’s *automatized* mode of action selection (based on composite actions), which is currently missing.

**Instrumental Value.** Drescher (1991) described instrumental value as a transient, context-sensitive, and goal-specific measure of an item’s utility. Items have instrumental value because they *currently* facilitate the acquisition of other things of value—with respect to a specifically foreseen chain of events (Drescher, 1991, p. 63).

Instrumental value could be assigned by LIDA's Action Selection module to component behaviors in a behavior stream. Selectable behaviors that lead to a behavior stream's goal state can be given instrumental value based on their proximity to that goal state. The amount of instrumental value those behaviors receive could be based, in part, on the desirability of the results that behavior stream's composite action is intended to effect (i.e., within a containing scheme).

**Templatized Schemes.** The LIDA conceptual model supports *templatized* schemes that contain one or more unbound variables in their contexts, actions, and results. For example, the templatized scheme **IN\_FRONT\_OF(OBJECT=?)/ PICKUP(OBJECT=?)/ HOLDING(OBJECT=?)** contains an unbound (unspecified) **OBJECT** variable. Templatized schemes represent a *family* of schemes from which specific instances can be created through the process of "instantiation." During instantiation, Procedural Memory creates an instance of a templatized scheme with its variables bound to content in a conscious broadcast. In the example above, the templatized scheme's **OBJECT** variable could be bound (during instantiation) to a **COFFEE\_MUG** node in a global broadcast, resulting in the instantiated scheme (i.e., behavior) **IN\_FRONT\_OF(OBJECT=COFFEE\_MUG)/ PICKUP(OBJECT=COFFEE\_MUG)/ HOLDING(OBJECT=COFFEE\_MUG)**.

Templatized schemes are powerful additions to the LIDA conceptual model that have no direct counterparts in Drescher's schema mechanism. Unfortunately, LIDA's computational implementation of templatized schemes and their instantiation operation is still poorly understood. While the learning of templatized schemes will likely require the development of a generalization process that learns templates from spin-offs, there is no reason to suspect that such a generalization process will be incompatible with the implementation outlined here. To the

contrary, it is likely that this generalization process could be largely agnostic of the processes responsible for learning individual, non-templated, schemes (e.g., marginal attribution). And it will almost certainly depend on processes external to Procedural Memory—for example, concept learning and conceptual generalization.

### ***Procedural Memory***

This section details my schema-mechanism-based implementation of LIDA's Procedural Memory module. It combines details from the classical schema mechanism with the modifications and enhancements discussed in the previous section (see Reconciling LIDA with the Schema Mechanism). This implementation preserves many aspects of Drescher's schema mechanism without modification; therefore, readers may wish to review the background section on Drescher's schema mechanism (provided earlier in this chapter).

**Basic Components.** Schemes are Procedural Memory's primary data structure. They have three main components: a context, an action, and a result. Traditionally, LIDA's contexts and results were encoded as node structures (see Chapter 4); however, this implementation encodes them as *sets of node assertions*. A node assertion stipulates the presence or absence of an *amodal* node in LIDA's global broadcast.

In addition to contexts, actions, and results, schemes also have base-level activations, extended contexts, and extended results. Extended contexts and extended results support instructionist procedural learning via marginal attribution.

**Instantiation.** A scheme's context can contain zero or more node assertions. For each global broadcast Procedural Memory receives, it compares the contents<sup>19</sup> of that broadcast against the node assertions in its schemes' contexts. A scheme's context is said to be *satisfied* when its node assertions are satisfied.<sup>20</sup> Such schemes are situationally relevant, and they are typically instantiated as *behaviors*. Procedural Memory then sends these behaviors to LIDA's Action Selection module to compete for selection and execution.

**Primitive and Composite Actions.** A scheme's action can be a primitive (built-in) action, or a composite (learned) action implemented by a *behavior stream*. Primitive actions are typically hard-wired to motor plans in LIDA's Sensory Motor Memory module. Composite actions are created when Procedural Memory identifies *chains of schemes* that lead to specific *goal states*. A collection of instantiated chains of schemes is referred to as a behavior stream.

Procedural Memory creates a new bare scheme (action-only scheme) for each learned composite action. Spin-off schemes can then be created containing those composite actions (via instructionist procedural learning).

**Instructionist Procedural Learning.** Procedural Memory must be initialized with a set of *bare schemes* (action-only schemes) that encode an agent's built-in actions. Bare schemes are

---

<sup>19</sup> This implementation of Procedural Memory ignores modal nodes, feeling nodes, sensory representations, and sensory content in the global broadcast. Schemes' contexts and results are only composed from amodal node assertions, and schemes are only activated/instantiated by the presence or absence of amodal nodes (in the global broadcast).

<sup>20</sup> This differs from LIDA's conceptual implementation of Procedural Memory, which uses structural matching to activate its schemes. Schemes are given current activation in proportion to the degree of semantic similarity between the contents of a global broadcast and a scheme's context. A scheme's result can also factor into this comparison (e.g., if the conscious broadcast contains an "option" to act).

typically hard-wired to built-in motor plans (in LIDA's Sensory Motor System) that fulfill the execution of those primitive actions. From this set of bare schemes, instructionist procedural learning (based on marginal attribution) can begin to construct *result spin-off* schemes based on the statistically significant consequences of an agent's actions. These spin-off schemes can then be used to construct *context spin-offs*—based on the incremental addition of amodal nodes that are statistically correlated with the successful execution of those schemes. This process of incremental refinement continues, creating new context spin-offs from previous schemes. Schemes with compound contexts (i.e., contexts containing multiple node assertions) enable the learning of result spin-offs containing those compound contexts, thus encouraging the identification of chains of schemes (see Background: The Schema Mechanism, Procedural Learning). Unlike context spin-offs, marginal attribution does not create result spin-offs incrementally.<sup>21</sup>

New amodal nodes can be learned using the grounded perceptual/conceptual learning mechanisms outlined in Chapter 5. These nodes function like *learnable* primitive items.<sup>22</sup> Whenever Procedural Memory encounters previously unseen amodal nodes in a global (conscious) broadcast, it will expand its extended contexts and extended results to include a new slot for each such amodal node. Procedural Memory will then begin tabulating correlation statistics for these new nodes.

---

<sup>21</sup> This is a computational optimization (see Drescher, 1991, pp. 78–79).

<sup>22</sup> The synthetic items used by the classical schema mechanism to support concept learning are not used in this implementation.

**Selectionist Procedural Learning.** Procedural Memory updates the *base-level activations* associated with its schemes following the execution of selected behaviors. In the traditional implementation of Procedural Memory, only a *single* scheme is updated per executed action—specifically, the scheme corresponding to a selected behavior. If the consequences of that selected behavior’s execution match the behavior’s predicted result, then its corresponding scheme’s base-level activation is increased. Procedural Memory makes this determination based on the contents of subsequent conscious broadcasts.

A more efficient way to perform selectionist procedural learning is to update *all* schemes that (1) share the same action as the selected behavior and (2) were instantiated as behaviors during the same cognitive cycle (as the selected behavior). Since a scheme is only instantiated when its context is satisfied<sup>23</sup>, its results can be compared to subsequent conscious broadcasts *as if it were the selected behavior*. This enhancement was inspired by Drescher’s schema mechanism<sup>24</sup>.

### ***Action Selection***

The standard implementation of LIDA’s Action Selection module (see Negatu & Franklin, 2002) has been described as an enhanced version of Maes’s (1989) behavior net<sup>25</sup>. LIDA’s behavior net selects *at most* one behavior per cognitive cycle based on the activation/inhibition-style

---

<sup>23</sup> This is a property of the current implementation that does not hold in the LIDA conceptual model. (The LIDA conceptual model supports the instantiation of schemes based on a *partial* context match.)

<sup>24</sup> Instantiated behaviors that share the same action as the selected behavior are analogous to what Drescher referred to as “implicitly activated schemas” (Drescher, 1991, p. 54). The classical schema mechanism updated the reliabilities of all such schemas following action execution.

<sup>25</sup> Maes’s behavior network is alternately referred to as the Agent Network Architecture (ANA; Maes, 1991).

dynamics between its behaviors. Activation spreads between behaviors over “successor/predecessor” links, and activation can be inhibited via “conflictor links” (see Maes, 1989). This behavior net-based implementation of Action Selection was also present in LIDA’s predecessor IDA (Intelligent Distribution Agent, see Franklin, 2003).

The schema-mechanism-based Action Selection implementation presented here is an *alternative* to the standard implementation, which is also consistent with LIDA’s conceptual model and its commitments. In particular, the proposed Action Selection implementation chooses at most one behavior per cognitive cycle from a set of relevant behaviors (instantiated schemes); it is compatible with all four modes of LIDA’s Action Selection; and it uses similar selection criteria, including those based on the *reliability* (base-level activation), *desirability* (total incentive salience of expected results), *applicability* (context satisfaction), and the *consequences* (e.g., instrumental value) of those behaviors.

I made the decision to use a schema-mechanism-based implementation of Action Selection here because instructionist procedural learning is extremely difficult to reconcile with Maes’s behavior net. IDA (LIDA’s predecessor) did not support learning; therefore, it was easier to coordinate the operations of Procedural Memory and Action Selection, even though they were not expressly designed to work together. While it may be possible to implement a procedural learning algorithm that works with Maes’s behavior net, it will require a Herculean effort that merits its own dissertation. That is well beyond the scope of this work, moreover, the benefits of doing so are unclear.

As a final note, there are several additional benefits to the implementation proposed here beyond its direct compatibility with the instructionist learning of schemes. First, the proposed

implementation introduces psychologically plausible exploratory criteria into the action selection process based on *hysteresis* and *habituation* (see Drescher, 1991, sec. 3.4.2). This trade-off between exploration and goal-directed behavior has never been explored in the LIDA literature. Second, the proposed schema-mechanism-based implementation suggests a possible implementation for LIDA's automatized mode of action selection, which has never been implemented<sup>26</sup>. Finally, the proposed implementation is highly extensible, easily supporting new selection criteria or the weighting of existing criteria according to their relative selection importance.

**Internal and External Behaviors.** The LIDA conceptual model includes both external and internal actions<sup>27</sup>. More generally, I propose that LIDA's Action Selection module should support the selection of behaviors for external or internal execution. The external execution of behaviors is fulfilled by LIDA's Sensory Motor System (see Chapter 4), which typically results in changes to an agent's (external) environment. Whereas a behavior's internal execution is carried out by structure building codelets and usually results in changes within LIDA's Current Situational Model (i.e., an agent's internal or mental environment).

---

<sup>26</sup> While out of the scope of this thesis, a computational implementation of automatized action selection could be based on the schema mechanism's composite actions, which are directly analogous to behavior streams. The chains of schemes within a composite action's controller could be selectively reinforced and decayed based on their execution history and the reliability of the component schemes. If a single chain of schemes within a given composite action's controller becomes the dominant chain (where all other chains have decayed away), then it becomes an automatized action. The selection of component schemes within such an automatized action could proceed in exactly the same way as the selection of components of a pending (composite action) scheme in the current implementation.

<sup>27</sup> Historically, LIDA's internal actions have been primarily used to facilitate its volitional mode of action selection, which is based on William James's ideomotor theory. Specifically, an internal action has been used to start a timer in LIDA's Current Situational Model following the selection of a deliberation behavior. This work greatly expands on that previous use case. In particular, I argue that most behaviors can be internally executed. And that the consequence of their internal executions is the mental simulation of those behaviors' expected results in LIDA's Current Situational Model.



There are considerable parallels between LIDA's action phase and Jeannerod's (1994, 1995, 2001) theory of motor cognition (see Chapter 2, Motor Cognition). In particular, when a LIDA agent's behaviors are selected for external execution, they function like Jeannerod's "overt" actions, and those selected for internal execution function like Jeannerod's "covert" actions. Furthermore, the "covert stage" representations hypothesized by Jeannerod (2001) are analogous to LIDA's instantiated schemes, which are the representational precursors of all of LIDA's selected behaviors. Finally, LIDA's internally executed behaviors (cf. Jeannerod's covert actions) support its implementation of *action-based mental imagery* (see Chapter 7), which is analogous to Jeannerod's hypothesis that covert actions enable motor imagery. Figure 19 shows these parallels between LIDA and Jeannerod's theory of motor cognition in the context of LIDA's cognitive cycle.

Drescher's schema mechanism does not include the equivalent of internal actions or mental simulations; therefore, an enhancement is needed to support this functionality in the schema-mechanism-based implementation described in this chapter. Specifically, whenever LIDA's Action Selection module chooses a behavior (instantiated scheme) for execution, it must also decide whether to execute that behavior internally or externally.

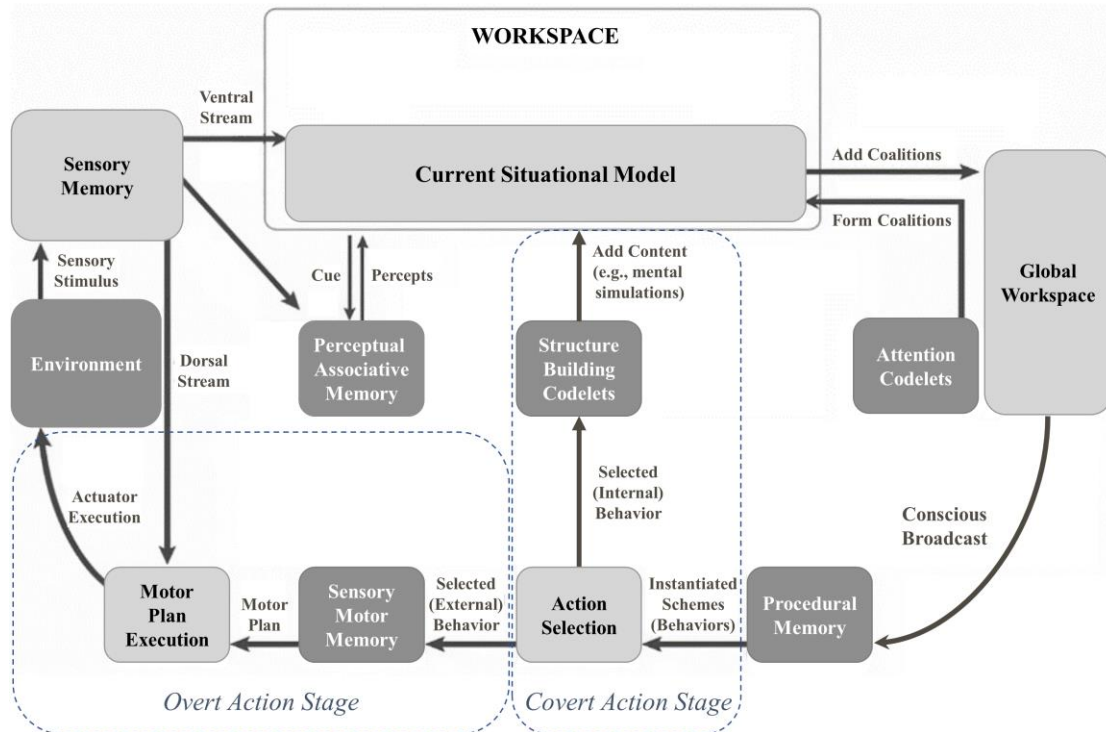


Figure 19. Jeannerod’s covert and overt action stages in LIDA. External (overt) behaviors are executed via LIDA’s Sensory Motor System (i.e., Sensory Motor Memory and Motor Plan Execution). Internal (covert) behaviors are executed by structure building codelets (e.g., simulator SBCs), resulting in action-based mental simulations. All overtly executed behaviors have an initial covert stage characterized by the instantiation of a set of schemes.

In some cases, this decision is trivial since not all behaviors are capable of being executed both internally and externally. For example, it’s typically impossible to internally execute (i.e., mentally simulate) behaviors based on bare (action-only) schemes because those behaviors have no associated expected results (i.e., they make no predictions). In other cases, an agent’s situational context may dictate that a behavior *must* be executed internally. For example, when an agent is deliberating about a distal intention (i.e., options for achieving goals for which external “overt” behavior cannot be immediately executed), it may engage in a form of “mental time travel” that precludes the external execution of those behaviors. An example of this occurs when one is deliberating on what route to take when driving to work. One does not physically take those potential routes to make this decision, but rather, mentally traverses those path options

and weighs the pros and cons of each. Similarly, agents may mentally manipulate objects that are not physically present in their environment (e.g., if an agent were playing a game of blindfolded chess). In these cases, agents engage in a form of “mental telekinesis” that requires internally executed behaviors.

Nevertheless, in many cases, behaviors can be executed either internally or externally, and LIDA’s Action Selection module must determine which is *more* appropriate or advantageous based on an agent’s current situation and the expected consequences of those actions. Numerous factors can influence this decision, including (but not limited to) the *time sensitivity* of a situation, perceived situational *risk*, perceived situational *uncertainty*, and the estimated *reliability of an agent’s internal model* (with respect to the agent’s current situation).

Naturalistic environments often place real-time demands on agents that preclude “offline” activities such as mental simulation. That is, time-critical circumstances (e.g., predator/prey interactions) may force agents to immediately execute “overt” re-active behaviors (lest they be eaten or starve). In LIDA, the most extreme example of this would be an *alarm* situation (see Chapter 4, Modes of Action Selection), which would necessitate that a reactive behavior is immediately selected for external execution, even if that behavior could have been executed internally. In other words, alarms are generally incompatible with the selection of behaviors for internal execution.

While it is certainly true that time pressure can be a powerful motivator for external “overt” behavior, it is also true that, given the opportunity, natural agents (e.g., humans) often “think about” their actions before externally executing them. They may assess their situations before acting, consciously weighing the pros and cons of potential actions, and develop action

plans that extend into places and times that are not externally perceivable. Many of these activities require internally executed behaviors (e.g., action-based mental simulations). Therefore, while it is clearly the case that time sensitivity influences whether Action Selection chooses a behavior for external or internal execution, it is not the only factor. For instance, it is typically ill-advised to rush to action in risky situations when there is time for additional contemplation. Similarly, if an agent is uncertain of their current situation, additional speculation and (internal or external) exploratory actions may be warranted. And, in general, any time an agent enters into a deliberative or volitional mode of action selection, internally executed behaviors, such as action-based mental simulations, are likely to occur.

As a final note, it is often the case that agents may develop, or have innate or built-in, proclivities that bias their action selection towards internal or external execution. More impulsive or reactive agents may favor externally executed behaviors, while more introspective, and risk-averse, deliberative agents may tend towards additional, preliminary contemplation (i.e., action-based mental simulation via internally executed actions).

## **Evaluation**

A software implementation<sup>28</sup> was developed based on the enhanced, LIDA-compatible, version of Drescher's schema mechanism described in this chapter. In order to evaluate this implementation, a modified  $k$ -armed bandit environment (explained below) was implemented along with a set of software agent that operated within that environment. An overview of this

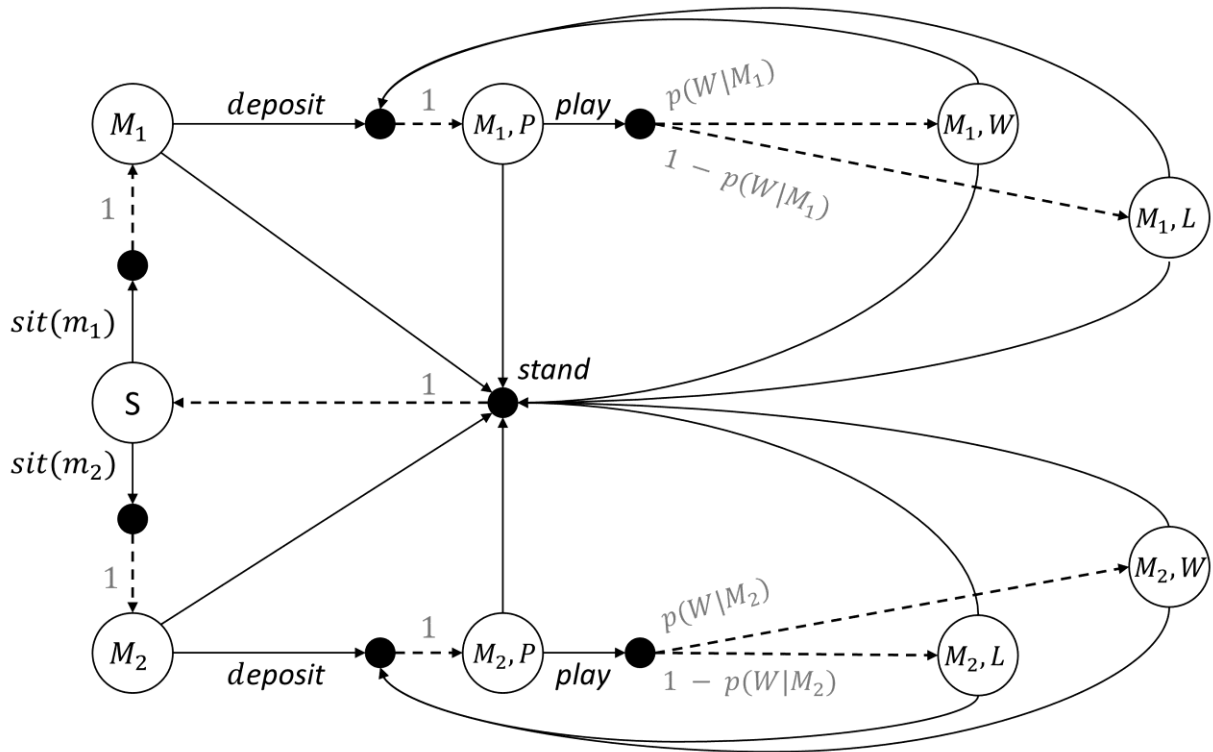
---

<sup>28</sup> <https://github.com/skugele/schema-mechanism>.

agent, its environment, and an analysis of the generated experimental results appear in the subsections that follow.

### ***Environment***

A standard  $k$ -armed bandit environment (see Sutton & Barto, 2018, sec. 2.1) contains  $k$  “levers” that an agent can pull. Each lever has a fixed probability of winning that an agent must infer by observing the environmental consequences of pulling those levers. If the agent pulls a lever and wins, it receives some payout (e.g., a monetary reward). If the agent pulls a lever and loses, it receives nothing. Agents operating in this environment are typically given a fixed budget (e.g., a limited number of lever pulls), and an agent’s objective is to maximize its profit by concentrating its actions (pulls) on levers with the highest expected value (cumulative payout).



*\*all actions that are not depicted result in no change to state*

<b>State Abbreviations</b>	
<b>S</b> : player standing	<b>W</b> : player wins
<b><math>M_i</math></b> : player seated @ machine $i$	<b>L</b> : player loses
<b>P</b> : player paid to play	

Figure 20. State-transition diagram for a modified,  $k$ -armed bandit environment. In this case,  $k = 2$ . White nodes depict environmental states. Black nodes depict actions. Links going from white (state) nodes into black (action) nodes indicate that an action is taken from that state. Links leaving action nodes are labelled with the probabilities of transitioning to various environmental states if the agent were to execute those actions from those states.

One shortcoming of this standard  $k$ -armed bandit environment is that the relationship between an agent's actions and their consequences are too simplistic to adequately test many fundamental agential capabilities. For example, each action (lever pull) is completely independent of the ones that preceded it, and only a single action is required to achieve something of value. More realistic environments typically require a series of coordinated actions to acquire something of value, and executing the wrong actions—or the right actions at the wrong time—can undermine an agent's progress towards its goals. In particular, the standard

$k$ -armed bandit environment does not test whether an agent can assign “partial credit” to actions that facilitate or enable the acquisition of something of value. This additional environmental challenge is often referred to as a *credit assignment problem* (see Minsky, 1961).

To address this shortcoming, a slightly more complicated  $k$ -armed bandit environment was developed. In this environment, an agent must first “sit” at a machine and “deposit” money into it before it can “play” that machine (i.e., pull its lever). If after playing a machine, an agent wishes to “play” the same machine again, it must once again “deposit” money into that machine before pulling its lever. If, on the other hand, the agent wishes to “play” a different machine, it must first “stand,” and then “sit” at the new machine. An agent may choose to “stand” at any time; however, it will lose any money it has deposited if it stands up before “playing” the machine containing the deposited money. Figure 20 depicts this modified  $k$ -armed bandit environment (with  $k = 2$ ) as a state transition diagram.

As a final note, the  $k$ -armed bandit environment used here for evaluation purposes was configured so that the “deposit” action cost an agent 1 credit<sup>29</sup>, while “winning” rewarded an agent with 2 credits.

### ***Agent***

A set of agents was implemented based on the modified schema mechanism described in this chapter. These implementations were focused primarily on Procedural Memory and Action

---

<sup>29</sup> This cost was only incurred when the agent executed the “deposit” action while sitting at a machine that did not already contain deposited money (that is, when executed in states  $M_i$ , states  $M_i, W$ , states  $M_i, L$ ).

Selection modules; however, a simple implementation of Perceptual Associative Memory was also required.

Each agent was initialized with  $k + 3$  bare schemes, where  $k$  was the number of machines in a  $k$ -armed bandit environment. This included bare schemes for `[/stand/]`, `[/play/]`, `[/deposit/]`, as well as  $k$  additional bare schemes for sitting at each machine—`[/sit( $M_i$ )/]`, for  $1 \leq i \leq k$ . Three feeling nodes were used to specify an agent’s built-in motivations associated with winning, losing, and depositing money in a machine. The event of “winning” produced a +1.0 affective valence, the event of “losing” produced a –1.0 affective valence, and the event of “depositing money” produced a –0.5 affective valence.

An amodal node was generated for each state element whenever it occurred in the environment. For example, if an agent were in the state  $M_1, W$ , this would generate two amodal nodes—one activated by  $M_1$  (the event of being seated at the first machine) and another by  $W$  (the event of having just won). Note that the composite event of  $M_1, W$  would also correspond to its own amodal node.

All nodes had the potential to accrue positive or negative base-level incentive salience based on their temporal proximity to states that generated affective valence. Base-level incentive salience updates were calculated using a temporal difference (TD) learning algorithm that used “replacing” eligibility traces (see Singh & Sutton, 1996). Current incentive salience and incentive salience links were not implemented.



Action Selection was based on a weighted<sup>30</sup> combination of *goal-directed* and *exploratory* evaluation criteria. Goal-directed evaluation criteria included: (1) the incentive saliences of schemes’ results (i.e., their desirability), (2) the schemes’ base-level activations (i.e., their reliability), (3) the instrumental values of schemes’ results (i.e., whether schemes help achieve the agent’s currently selected goal state), and (4) “pending focus” (i.e., a selected behavior stream’s—pending scheme’s—component schemes are given additional selection importance). Exploratory criteria included: (1) habituation (i.e., recently or frequently selected schemes have decreased importance; see Drescher, 1991, sec. 3.4.2), and (2) “temperature-based” randomized exploration<sup>31</sup>.

## ***Results***

Two sets of experiments were conducted to test the implementation’s capabilities. Both used the modified  $k$ -armed bandit environment described earlier in this section, but with different configurations. The agents’ runtime parameters (see Table 7 in the Appendix) were tuned using Optuna (Akiba et al., 2019) based on an eight machine ( $k = 8$ ) environment with randomly generated win probabilities. These parameter values remained fixed throughout all evaluation trials (regardless of  $k$ ).

---

<sup>30</sup> The weights associated with these goal-directed and exploratory criteria varied cyclically, so that an agent’s actions oscillated between being more goal-directed or more exploratory (see Drescher, 1991, sec. 3.4.2).

<sup>31</sup> A “temperature” parameter ( $0.0 \leq \epsilon \leq 1.0$ ) was used to control the probability that Action Selection would randomly choose a single scheme to receive a bonus to its selection importance during a selection event. Initially,  $\epsilon$  was set close to 1.0 to encourage exploration. However, its value was decreased over time, reducing the degree of random exploration. This idea is similar to the temperature used in simulated annealing (Kirkpatrick et al., 1983) or the epsilon parameter used in reinforcement learning’s epsilon-greedy policies (see Sutton & Barto, 2018, sec. 5.4).

**Experiment 1.** In the first set of experiments, modified  $k$ -armed bandit environments were configured where the *range* (min = 0.0; max = 1.0) and *average* (0.5) over their machine’s win probabilities remained the same, but the number of machines varied ( $k = \{2, 4, 8, 16, 32, 64\}$ ). Note that the differences between individual machine’s win probabilities decreased as the number of machines ( $k$ ) increased (see Table 2).

This set of experiments was intended to test the implementation’s ability to prefer the machines with the highest win probabilities. Agents were given a fixed budget of 25,000 actions; 30 trials were executed for each environment configuration, using different random seeds. The results are summarized in Figure 21.

Table 2. Percentage of trials in which agents focused on the best machine (experiment 1). The approximate difference between best and 2nd best machines per  $k$  is also shown<sup>32</sup>.

$k$	<b>Trials Focused on Best</b>	<b>Approximate Difference in Win Probabilities</b>
<b>2</b>	100.00%	1.00
<b>4</b>	96.66%	0.25
<b>8</b>	63.33%	0.08
<b>16</b>	53.33%	0.04
<b>32</b>	23.33%	0.02
<b>64</b>	16.66%	0.01

---

<sup>32</sup> For  $k > 4$ , the differences in the win probabilities between the best two machines and worst two machines were *half* the difference between other consecutive machines. For example, the complete win probabilities for  $k = 8$  were [0.0, 0.083, 0.25, 0.417, 0.583, 0.75, 0.917, 1.0]. The best (and worst) two machines differed in their probabilities by 0.083, while other consecutive probabilities differed by 0.167.

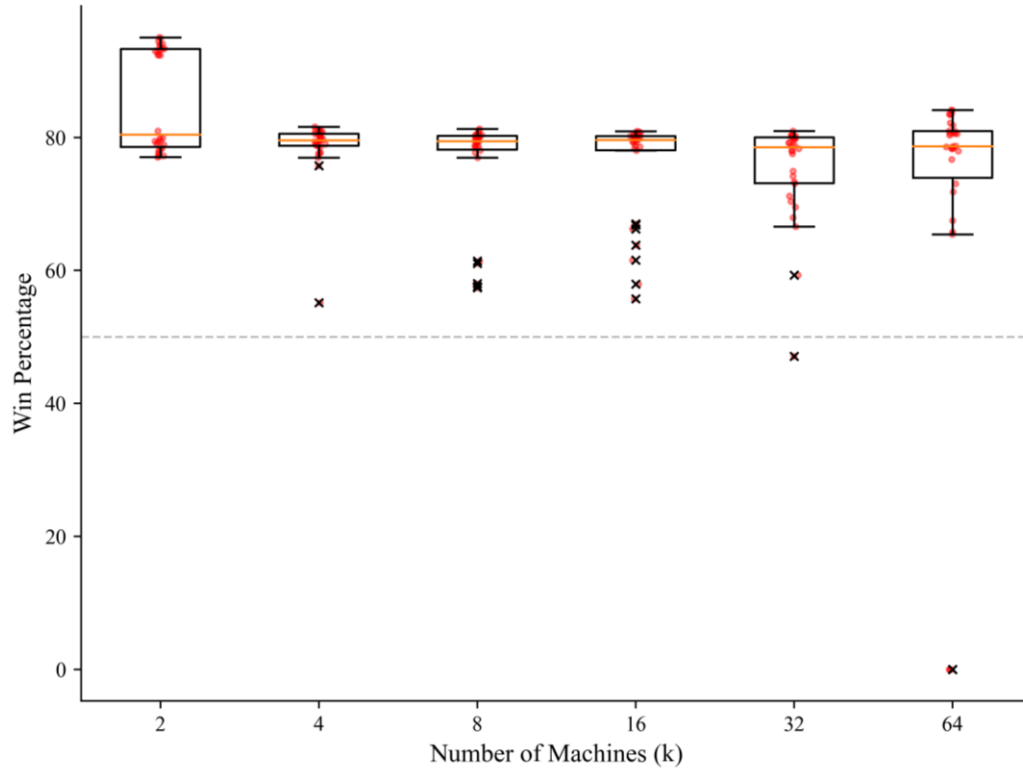


Figure 21. Boxplot of experiment 1’s results. The  $k$ -armed bandit environment was configured with a variable number of machines,  $k = \{2, 4, 8, 16, 32, 64\}$ , but the same min (0.0), max (1.0), and average (0.5) machine win probabilities. 30 trials were run for each value of  $k$ . Over all experiments (i.e., for all  $k$  values), agents achieved a median win percentage of approximately 80%. The optimal win percentage was 100% if agents knew the best machine to play in advance. The agents’ median win percentage per trial is shown using a solid, orange horizontal line. A gray, dashed, horizontal line shows the expected average win percentage given completely random machine choices.

Agents achieved a median win percentage—over all trials regardless of  $k$ —of around 80%. A win percentage of 50% would have been at chance and 100% would have been optimal—though this is unattainable in practice. In the majority of trials with  $k \leq 16$ , agents preferred *the best* machine over other machines; however, this preference dropped considerably as the number of machines increased and the differences in probabilities decreased (see Table 2). Doubling the agents’ budgets to 50,000 only slightly increased the percentage of trials in which agents selected the best machine most often (from 23.33% to 31.25% for  $k = 32$ ). Therefore, the

limiting factor may be the relative closeness in the expected payouts between the best machines as  $k$  increases (differing by less than 2% for  $k = 32$ ).

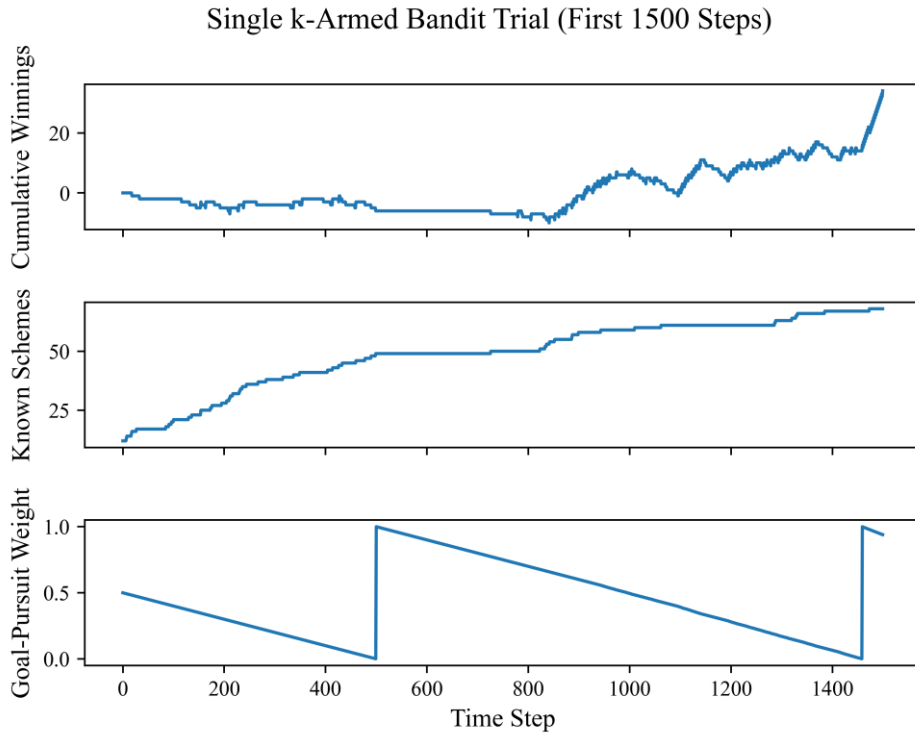


Figure 22. Plots of the first 1500 steps from a single  $k$ -armed bandit trial ( $k = 8$ ). This trial was chosen because its final cumulative winnings (after 25000 steps) was close to the median value from the 30 earlier  $k$ -armed bandit trials (where  $k = 8$ ) from experiment 1. The first sub-plot shows the agent’s cumulative winnings. (Recall that the agent must spend 1 credit to play a machine, and it is awarded 2 credits on a winning pull). The second sub-plot shows the number of known (learned + innate) schemes. The final sub-plot shows Action Selection’s goal-pursuit weight as a function of time. (Recall that the balance between goal-pursuit and exploration is cyclically updated.)

Figure 22 shows the first 1500 steps of a “typical” trial when  $k = 8$ . This trial was chosen because its final cumulative winnings (after 25,000 steps) were closest to the median value. Sub-plots show the agent’s cumulative winnings, total number of schemes (built-in and learned), and the goal-pursuit weight<sup>33</sup> as functions of time. Notice that starting at time step 1459

---

<sup>33</sup> The weight given to exploratory scheme selection criteria is 1.0 minus the goal-pursuit weight.

this agent began to single-mindedly prefer the best machine—playing it almost exclusively. This resulted in an overall winning trend that continued for the remainder of the trial (with small oscillations when the weight of the exploratory selection criteria strongly dominated that of the goal-directed selection criteria).

After 25,000 steps, this agent had learned 69 schemes in addition to its 11 built-in schemes (see Table 8 in the Appendix). The majority of these were learned by step 1500. The amodal nodes learned and their base-level incentive salience values are shown in Table 9 of the Appendix. The most important thing to note is that the numerical ordering of the base-level incentive saliences associated with the events of sitting at a machine ( $M_i$ ) and those for sitting at a machine with money deposited in that machine ( $M_i, P$ ) are correctly ordered based on the win probabilities of those machines. That is, the agent was capable of correctly evaluating the “desirability” of sitting at, and depositing money in, those machines.

**Experiment 2.** In a second set of experiments, modified  $k$ -armed bandit environments were once again configured with a variable number of machines, ranging from 2 to 64 ( $k = \{2, 4, 8, 16, 32, 64\}$ ); however, this time only a single machine was given a high win probability (0.9). All other machines had a low win probability (0.1). As a result, the average probability of winning decreased as  $k$  increased. This set of experiments was intended to test an agent’s ability to find and exploit the environment’s single “good” machine when the payouts became sparser. As in experiment 1, agents were given a fixed budget of 25,000 actions, and 30 trials were executed using different random seeds for each environment configuration.

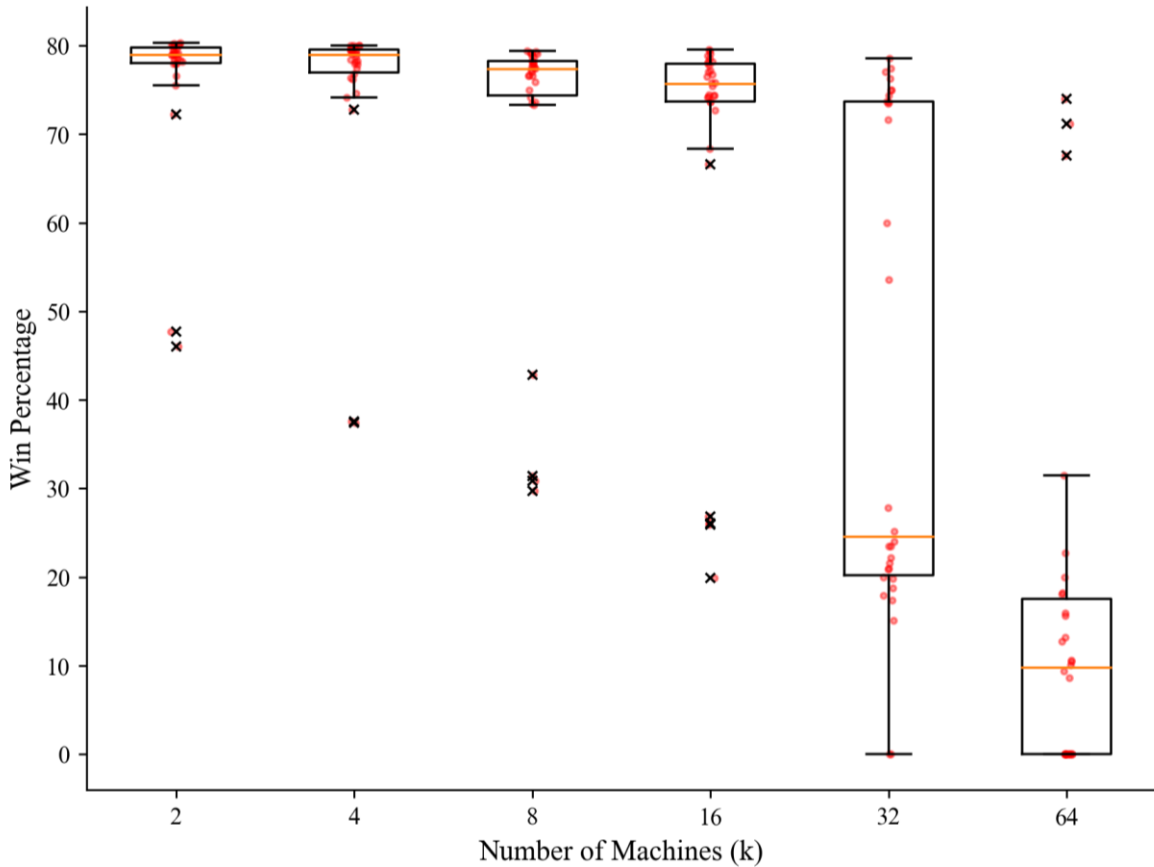


Figure 23. Boxplot of experiment 2's results. The  $k$ -armed bandit environment was configured with a variable number of machines,  $k = \{2, 4, 8, 16, 32, 64\}$ . Only a single machine had a high probability of winning (0.9). All other machines had a low probability of winning (0.1). As a result, the chance win percentage decreased with the number of machines: 50% ( $k = 2$ ); 30% ( $k = 4$ ); 20% ( $k = 8$ ); 15% ( $k = 16$ ); 12.5% ( $k = 32$ ); 11.25% ( $k = 64$ ). 30 trials were conducted for each value of  $k$ . Over all experiments (i.e., for all  $k$  values), the optimal win percentage was 90%; however, achieving this would require that agents knew the best machine to play in advance. The median win percentage per trial is shown using a solid, orange horizontal line.

The results of experiment 2 are summarized in Figure 23. A win percentage of 90% would have been optimal. The chance of randomly winning decreased with increasing  $k$ : 50% ( $k = 2$ ); 30% ( $k = 4$ ); 20% ( $k = 8$ ); 15% ( $k = 16$ ); 12.5% ( $k = 32$ ); 11.25% ( $k = 64$ ).

Figure 23 shows that agents were generally able to identify and exploit the single high probability machine when  $k \leq 32$ ; however, variability in the trials increased dramatically for  $k > 16$ , and the median win probability for the  $k = 64$  trials was *at chance*. Doubling the

agents' budgets to 50,000 did little to correct this issue. This suggests that the agent's exploratory selection criteria need to be adjusted for larger, sparser environments. In particular, the  $\epsilon$  temperature-based exploratory criteria reached its minimum value at step 36,887—effectively eliminating most exploration—since the influence of the habituation-based criterion appears to be too weak to promote significant exploratory actions in its current form.

## **Discussion**

In this chapter, I described a new conceptual and computational implementation of LIDA's Procedural Memory and Action Selection modules. This implementation is based on a heavily modified and enhanced version of Drescher's schema mechanism. The experimental results suggest that, starting from only a modicum of knowledge (i.e., primitive actions and built-in feeling nodes), this implementation can construct an internal model of its environmental interactions. Moreover, it can use that internal model to select goal-directed or exploratory behaviors. While these results have only been demonstrated on a simple stochastic environment, the results are promising.

This implementation significantly advances LIDA's instructionist procedural learning, provides a computational mechanism for behavior streams, *suggests* an implementation for automatized action selection, expands the use cases for internal actions to include *action-based* mental simulations, and proposes an enhancement to LIDA's Action Selection module that controls the selection of behaviors for external or internal execution. Furthermore, an explicit connection was made between Jeannerod's theory of motor cognition and LIDA's action phase. An additional benefit of this implementation is that its operations are transparent and easily interpreted. For example, it is straightforward to ascertain *why* an agent selected a particular

behavior, from the conscious content that resulted in the instantiation of a set of selectable behaviors, to the selection importance assigned to each.

The internal model learned by this implementation is the foundation of action-based mental simulation in LIDA. Furthermore, I contend that the selection of behaviors for internal execution is the backbone of simulation-based cognition—working in conjunction with multimodal perception (see Chapter 5) to implement imagistic, epistemic processes (see Chapter 7).



## Chapter 7

### Mental Imagery and Simulation-Based Epistemic Processes

perceptual experience falls at varying positions along a continuum between the extremes of pure stimulus and pure imagery... imagery is not simply a thing apart, an internal representation distinct from the scene before our eyes, but rather it is part-and-parcel of perception. (Albright, 2012, p. 235)

This chapter explores mental imagery and its implementation in LIDA. Using the foundations established in Chapters 5 and 6, I implement the four fundamental imagistic operations underpinning simulation-based cognition—image generation, image transformation, image inspection, and image maintenance (see Kosslyn, 1994; Kosslyn et al., 2006). Generative processes produce mental simulations; transformative processes alter the properties and parts of those simulations; introspective processes selectively orient attention towards situationally relevant properties and parts; and maintenance processes ensure that mental simulations persist long enough to be manipulated and perceived. Taken together, these operations—combined with multi-modal perception—provide a mechanism for generating new knowledge *without* rule-based symbolic manipulation.<sup>1</sup>

The chapter begins with background on mental imagery: its nature, its properties, and the four fundamental imagistic operations that are believed to support it. Following this, I describe

---

<sup>1</sup> Recall from Chapter 2 that classical symbolic cognitive theories hypothesize that natural cognitive systems employ epistemic processes that are primarily based on symbolic manipulations (e.g., see “the physical symbol system hypothesis,” Newell & Simon, 1976). In contrast, embodied, simulation-based theories of cognition contend that natural systems internally construct new knowledge through the coordinated activity of imagistic and perceptual processes.

Kosslyn's "proto-model" of visual imagery and perception (Kosslyn, 1994), which inspired portions of the LIDA-based implementations described here. Next, I detail my LIDA-based conceptual implementation of mental imagery and the design and partial computational implementation of a simulation-based LIDA agent. The chapter concludes with a comparison of my implementation to those that appear in several other cognitive architectures, as well as a brief discussion.

## **Mental Imagery**

Most humans report the ability to have sensory-like experiences in the absence of external stimuli. This has led to expressions such as "seeing with your mind's eye," "having a song stuck in your head," and "listening to your inner voice." Collectively, these phenomena are referred to in the literature as *mental images* (i.e., consciously accessed mental simulations). Indeed, many humans are capable of using mental imagery to create complex, multimodal, virtual scenes, and they do so both spontaneously (e.g., when daydreaming), or intentionally to facilitate a variety of cognitive processes (e.g., planning and mental rehearsal).

In its most basic form, mental imagery enables the reconstruction of prior experiences in a way that conveys sensory-like qualities on those remembrances. However, mental imagery's utility extends well beyond the simple reconstruction of past experiences. Mental images often function as *experience-based predictions* (see Moulton & Kosslyn, 2009) that allow us to anticipate what we might encounter, or imagine what we might have encountered if circumstances had been different (i.e., counterfactuals). These mental simulations are "epistemic devices" (Fisher, 2006) that generate (or make available) *new knowledge*. Moulton and Kosslyn (2009) stated,

[mental imagery] allows us to simulate reality at will, and, because of this, allows us to predict what we would experience in a specific situation or after we perform a specific action. This ability not only allows us to reconstruct the past, but also to anticipate what may occur in the near and distant future. (p. 1273)

In other words, mental imagery not only supports the recreation of prior experiences, but the productive elaboration, transformation, and recombination of those experiences as a way of understanding the world. This ability to generate and manipulate imagery-based simulations may underlie many higher-order cognitive processes such as *problem solving* (Clement, 1994; Y. Qin & Simon, 1992; Shaver et al., 1974), *mental rehearsal* (Driskell et al., 1994; Keller, 2012), and *language comprehension* (Bergen et al., 2007; Bergen & Chang, 2005; Zwaan, 2004, 2014; Zwaan & Taylor, 2006). Moreover, these conscious mental images are indicative of a more ubiquitous cognitive phenomena—mental simulation—that may directly support the majority of our cognitive processes, both overtly and covertly (Barsalou, 1999, 2016b; Grush, 2004; Hesslow, 2002, 2012).

### ***Properties of Mental Images***

Mental images can be described as having the following basic properties:

- consciously accessible
- analogous to stimuli from sensory modalities
- interacts with externally sourced (i.e., real) sensory representations
- rapid decay

Each of these properties is briefly discussed in the subsections that follow.

**Conscious Accessibility.** By definition, mental images are *consciously accessible*, and they are consciously experienced. This is not to say that preconscious precursors of mental images do not exist or that those precursors do not serve a functional role in cognitive systems. In fact, a major goal of this work is to elaborate on the preconscious precursors of mental imagery and to explicate their functional roles in support of cognitive tasks. However, the fact that mental images are consciously accessible is a tremendous boon. It affords us a glimpse into many aspects of mental simulations that we would be unable to discern otherwise (for example, through introspection and self-report).

Ideally, many of the properties attributed to mental images would also apply to their preconscious counterparts (mental simulations); however, there is a possibility that their properties may differ. For example, Barsalou (1999) suggested that unconscious (i.e., preconscious and never-conscious) simulations might violate certain constraints that are compulsory for their conscious counterparts. While a mental image of a cup must minimally have an orientation, a shape, and subtend some angle in the (internal) visual field, an unconscious mental simulation of a cup may relax these requirements, such that, one or more of these aspects may be “inactive” at any given time. Assuming this hypothesis is correct, it leads to many interesting questions and challenges in modeling such phenomena.

**Analogous to Stimuli from Sensory Modalities.** Mental images and their preconscious counterparts (i.e., mental simulations) appear to be based (at least in part) on depictive,

topographically<sup>2</sup> organized, non-symbolic representations (see Kosslyn et al., 2006, Chapter 4). And many of the same brain regions that support mental imagery also support the perception (Chen et al., 1998; Ganis et al., 2004; Klein et al., 2004; Kosslyn et al., 1995, 1997, 1999; Slotnick et al., 2012).

**Interaction between External and Internal Sensory Representations.** Based on psychological and neuroimaging studies, it appears that real and virtual sensory content can interact in various memory systems. In particular, it seems that imagination has the ability to either interfere with or prime perceptions of real phenomena, and vice versa. The earliest and most famous study demonstrating this interaction was by Perky (1910), and the results she described are now referred to as “the Perky effect” (Waller et al., 2012). Numerous studies have followed in Perky’s (1910) tradition and established a complex set of interactions between real and imagined stimuli (Craver-Lemley & Reeves, 1992; Farah, 1985, 1989; Pearson et al., 2008).

Albright (2012) contended that our perceptions are based on a mixture of external stimuli (from sensory organs) and internal stimuli (from mental simulations). The relative contribution of each is largely based on the strength and quality of those stimuli. These, in turn, may be based on the fidelity of our external sensory inputs (e.g., one’s visual acuity), and the relevance, reliability, and precision of our internal models of the world.

This idea makes a certain amount of intuitive sense. When the external environment provides you a clear and unambiguous sensory signal, it should be preferred over internally

---

<sup>2</sup> Topographical organization entails that cortical locations can be used to represent locations in space (e.g., points within the visual field) and that the distances between those cortical locations are roughly proportional to the distances between the locations they are intended to represent.

generated and more speculative representations. On the other hand, when external sensory stimuli are highly ambiguous or degraded, then internal predictions should take over to supply missing details. In between these extremes, real and virtual sensory content should collaborate to realize a coherent and plausible perceptual interpretation of one's current situation.

**Rapid Decay.** Without continual maintenance, mental images rapidly lose clarity and become inaccessible to conscious experience. Kosslyn et al. (2006) argued that this rapid fading of mental images is a consequence of shared sensory cortical regions (e.g., topographically organized areas of the primary visual cortex) that are used to represent both real and imagined sensory content. They further argue that this decay is necessary to prevent prior real and imagined sensory content from interfering with incoming sensory content.

### ***Fundamental Operations***

Kosslyn (1994) defined four fundamental cognitive operations that support mental imagery: generation, transformation, inspection, and maintenance. Each of these operations are described below.

**Image Generation.** Mental images can be formed in numerous ways. They can be generated *spontaneously* (i.e., through involuntary unconscious processes) or *intentionally* (i.e., through volitional or consciously mediated processes). Their generations may be *single-part* (constructed in a single operation) or *multi-part* (requiring sequential elaboration). And their content can be *sensory* (e.g., the simulated experience of a sensory event, such as the smell of garlic) or *motor* (e.g., the simulated execution of an action and its environmental consequences, such as turning a doorknob).

Spontaneous generation—also referred to as involuntary (Brewin et al., 2010) or implicit (Albright, 2012) imagery—is initiated “automatically” via unconscious cognitive processes. For example, spontaneous mental images can occur in response to cued memories or through the efforts of predictive processes. We engage in this form of mental simulation when we preconsciously “fill-in” an object’s missing details (for example, when part of an object is occluded), or when we conjure plausible sensory interpretations of nebulous shapes or indistinct noises. We may also experience spontaneous imagery during language comprehension (Bergen, 2012). Albright (2012) argued that spontaneous mental imagery is “fundamental and ubiquitous,” (p. 235) serving to augment noisy, ambiguous, or otherwise incomplete sensory data based on probable predictions about their sources. In the extreme, spontaneous imagery can also become intrusive and pathological, and is linked to numerous psychological disorders, such as post-traumatic stress disorder, social anxiety disorders, eating and body perception disorders, and schizophrenia (Brewin et al., 2010).

Intentional generation—also referred to as voluntary (Brewin et al., 2010), volitional (Kreiman et al., 2000) or explicit (Albright, 2012) imagery—requires conscious mediation and is often accompanied by an explicit intention to engage in imagistic thought. This type of mental imagery is “conjured on demand [in service of] specific cognitive or behavioral goals” (Albright, 2012, p. 234). Players of strategy board games (such as chess or Go) may engage in intentional mental imagery when they try to imagine what the board would look like if their pieces were arranged differently. Tasks that require the recognition of spatial or temporal relationships between objects (or their parts), or judgments about their sensory qualities, may also initiate volitional processes that utilize intentional mental imagery.

Image generation exhibits predictable timing effects based on the complexity of the generated mental images. More complex, multi-part images require more time to generate than less complex or single-part images. Consequentially, image generation is believed to be governed by sequential, rather than parallel, cognitive processes (Kosslyn et al., 1988). Kosslyn (1994, p. 292) proposed that multi-part, visual images could be formed by first generating a global (“skeletal”; see Kosslyn, 1980) image that depicts the overall extent of an object. Representations for individual object parts could then be sequentially activated to generate images of those parts at specific locations on the global topographical extent, as needed. The specific order of this sequential elaboration depends, at least in part, on image inspection (described later).

**Image Transformation.** Image transformations include any operation that starts from an existing mental simulation and manipulates it in some way to change its sensory or motor characteristics. Operations such as mental rotations, scanning, and zooming are frequently given examples. However, other visual and non-visual transformations, such as auditory pitch and tactile pressure, also apply.

Chronometric (reaction time) studies are frequently cited as evidence that brains perform image transformations, particularly in support of perceptual tasks. Participants in these studies are typically required to make same-difference judgments when shown two objects with varying orientations, sizes, etc. Their reaction times are then recorded. In most cases, these reaction times are described as a *linear* function of the degree of perturbation between the presented objects. For example, the time subjects require to make same-difference judgements about rotated 2D and 3D objects is a linear function of the angle of presentation (Cooper, 1975, 1976; Metzler &



Shepard, 1974; Shepard & Metzler, 1971): the greater the degree of rotation between presented objects, the greater a subject's reaction time. From these results, researchers have hypothesized that their subjects are performing a series of *incremental* mental transformations that enable them to perform these perceptual discrimination tasks.

Georgopoulos et al. (1989) generated further support for this hypothesis by directly recording the patterns of activity in the motor cortex of a rhesus monkey while it performed a task requiring physical rotations. This monkey was trained to move its arm "in a direction that was perpendicular to and counterclockwise from the direction of a target light that changed in position from trial to trial" (Georgopoulos et al., 1989, p. 234) Georgopoulos et al. observed a sequence of preparatory, counterclockwise, mental rotations in the monkey's "neuronal population vector" in the moments prior to the monkey moving its arm. They interpreted this as direct support for the mental rotation hypothesis.

Finally, Zacks (2008) made several additional observations about mental rotation based on a meta-analysis of 32 high-quality, neuroimaging (PET or fMRI) studies. His first observation was that there is substantial evidence supporting the hypothesis that mental rotation appears to be based on the sequential transformation of "analog spatial representations" rather than abstract, descriptive representations. A second observation was that motor simulations were often involved in performing these transformations, though the degree of involvement depended on the specifics of the subjects' tasks. For example, Kosslyn et al. (2001) found that PET showed *primary* motor cortex involvement when subjects imagined themselves rotating an object, but not when they imagined an external force rotating that object. However, the *premotor* cortex was involved in both cases. As a result, it appears that some motor processes are employed whenever

we imagine objects rotating, but that the processing remains at a relatively high level unless subjects imagine themselves physically manipulating those objects. Moreover, Zacks noted that there appears to be an interaction (either interference or facilitation) between mental rotation and manual rotation tasks, depending on the task configuration. For example, Wexler et al. (1998) found that physically rotating a joystick handle in the same direction as required by a simultaneous mental rotation task decreased reaction times and error rates, while physical rotation in opposite direction increased reaction times and error rates.

**Image Inspection.** Image inspection refers to the ability to *volitionally* explore, perceive, and attend to portions of mental images. For example, the “mental scanning” of visual images is a frequently used experimental paradigm for investigating the properties of image inspection. In one such experiment, Kosslyn et al. (1978) had subjects memorize the map of a fictional island that contained seven objects (hut, tree, rock, well, lake, sand, and grass) situated at various locations. Once subjects had demonstrated that they had learned the map (by accurately drawing it from memory), they were asked to mentally picture the entire map. They were then asked to focus on one of the objects depicted on the map. Following a brief pause, they were given a second named object, and asked to scan to that new location on the imagined map. Subjects were instructed to perform this scanning operation by imagining a small dot traversing the shortest, straight-line path between the two objects. They were also instructed to press a button when their mental dot had reached the destination object. Kosslyn et al. (1978) reported a linear increase in subjects’ reaction time with the map distance scanned. They hypothesized that subjects were utilizing a form of image transformation—“in the same class as mental rotation and size alternation” (Kosslyn et al., 1978, p. 60)—to perform this scanning operation.

Finke and Pinker (Finke & Pinker, 1982, 1983; Pinker et al., 1984) later conducted a series of experiments that addressed several criticisms of Kosslyn et al.'s (1978) experimental procedure. In particular, in Finke and Pinker's experiments, subjects were never instructed to conjure mental images, nor were they explicitly told to perform mental scanning. Instead, subjects were instructed to remember a pattern of four dots on a display. The dots were subsequently removed, and an arrow was displayed. Subjects were required to report whether the arrow pointed to any of the previously displayed dots, where the distance between the arrow and the dots varied between trials. Once again, reaction times were found to increase linearly with distance.

After completing these trials, subjects were asked to describe how they mentally performed these tasks. Most participants reported mentally scanning a remembered dot pattern, starting from the location of the arrows, and in the directions indicated by those arrows. Pinker et al. (1984) hypothesized that this scanning operation was performed by mentally shifting an *attentional locus* (Pinker et al., 1984, p. 216), in small increments, along an imagined visual field. Notably, this mechanism differs from the transformation-based scanning operation postulated by Kosslyn et al. (1978).

Borst et al. (2006) investigated the possibility that different cognitive processes were being used to perform mental scanning, depending on the specifics of the task. They did this by repeating the earlier experiments, subjecting each of their participants to *both* experiments, and analyzing their reaction times. Borst et al. hypothesized that if the same underlying cognitive processes were being used in both cases, then they would expect to see highly correlated reaction time/distance slopes between the two experiments for the same subjects. Conversely, if different

cognitive processes were being employed, they surmised that the slope correlations between experiments would be weaker (for a given participant). Not only were Borst et al. (2006) able to replicate the findings of the earlier experiments, but they observed no significant correlations between the reaction time slopes for individual participants. As a result, they concluded that (at least) two different cognitive processes were being used to perform mental scanning.

Image inspection also seems to engage some of the same neural areas as image generation (Kosslyn et al., 2004). This is not surprising, as inspection often drives the sequential elaboration of mental images, adding parts and properties as needed (see Kosslyn, 1994, Chapter 9). That said, the processes of image generation and inspection are distinct. For example, Kosslyn et al. (2004) found that numerous brain regions were independently (i.e., not jointly) activated during image generation and image inspection tasks.

All of this to say that image inspection is a complex process that seems to enlist cognitive mechanisms associated with many of the other imagistic operations, as well as attentional processes. Furthermore, it is well-established that image inspection and perception share cognitive processes and neural substrates (Ganis et al., 2004; Kosslyn et al., 1997; Slotnick et al., 2012).

**Image Maintenance.** Kosslyn stated that image maintenance “lies at the heart of the use of imagery in reasoning” (Kosslyn, 1994, p. 324). The rapid decay of mental images requires the ability to actively maintain mental images in working memory. While not as glamorous as the other functions, offline cognitive processes (e.g., reasoning and planning) require the retention of images in short-term memory, working memory, and sensory memory (iconic/echoic memory) long enough to manipulate and inspect them. Therefore, this capability may be a basic

prerequisite for any meaningful use of mental images in a cognitive system. Kosslyn (1994) argued that image maintenance involves the repeated activation of representations by selectively attending to portions of mental images, or through the repeated engagement of attentional processes that operate on “the same loci in the visual buffer” (Kosslyn, 1994, p. 325).

### **Kosslyn’s Proto-Model of Visual Perception and Mental Imagery**

This section describes Kosslyn’s (1994) “proto-model” for *visual* perception and mental imagery, which inspired portions of the LIDA-based implementation described later in this chapter. Kosslyn’s proto-model is comprised of seven components (i.e., subsystems, which are depicted in Figure 24). Each subsystem will be briefly described in the sections that follow, along with their relevant representations and processes. Kosslyn (1994) and Kosslyn et al. (2006) provide neurophysiological and experimental evidence supporting each of these subsystems. For brevity, those details will not be repeat here.

#### ***Visual Buffer***

The Visual Buffer corresponds to topographically organized areas of the occipital lobe that “depict” the geometry of a stimulus’s retinal (or mental) projection. The Visual Buffer functions as a *non-symbolic* spatial extent in which cortical regions are used to represent retinotopic extents. According to Kosslyn et al. (2006), the low-level properties of environmental stimuli (such as color, intensity, depth, and motion) can then be encoded at locations within the Visual Buffer using a *symbolic* (propositional) code. As such, the Visual Buffer is a *hybrid* (symbolic/non-symbolic) representation—a non-symbolic, topographically organized extent overlaid with symbolically encoded shape properties.

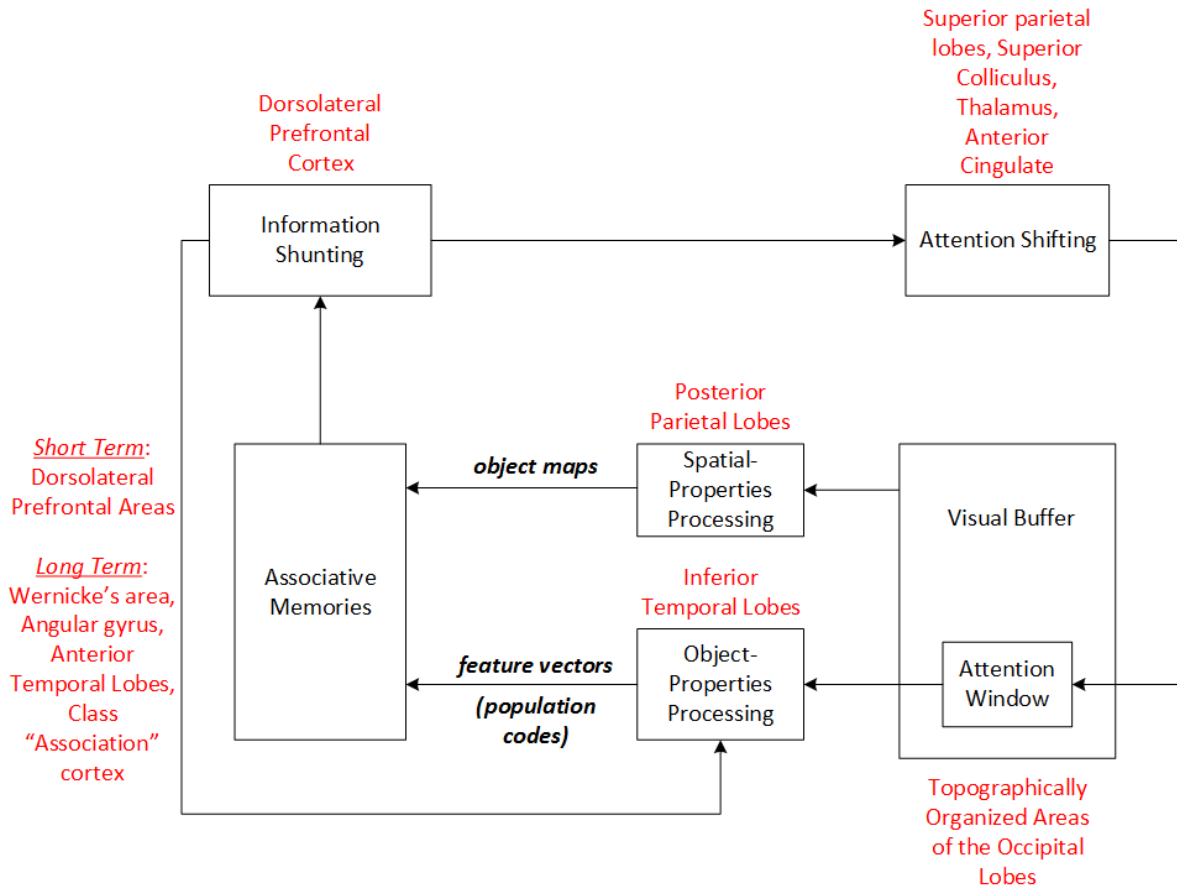


Figure 24. Kosslyn's proto-model of visual perception and mental imagery. The protomodel contains seven subsystems, which are depicted above. The brain regions believed to correspond to each subsystem are shown in red font next to their subsystems.

Kosslyn et al. (2006) characterized the Visual Buffer as a “canvas” upon which mental images could be depicted. Furthermore, they assumed that the proto-model’s other subsystems could operate on mental images in essentially the same way as they would externally sourced environmental stimuli.

### ***Attention Window***

The Attention Window supports the system’s ability to selectively focus on “contiguous sets of points” in the Visual Buffer (Kosslyn, 1994, p. 70). The focal region of the Attention Window can be shifted (e.g., spontaneously or volitionally) to center on different areas of the Visual

Buffer. This shifting might occur, for example, during image inspection. Kosslyn argued that the Attention Window supports both mental imagery and perception (Kosslyn, 1994, p. 101).

### ***Object-Properties Subsystem***

Kosslyn et al. (2006) asserted that object properties (e.g., color, texture, and shapes) are encoded in brains using *population codes* (or feature vectors; Fujita et al., 1992). They referred to the subsystem responsible for encoding these properties as the Object-Properties Subsystem. The population codes produced by this subsystem are based on sets of neurons (within the inferotemporal cortex) that selectively respond to different object-property dimensions. Different combinations of these neurons can be used to specify, and later recognize, an object's properties.

Kosslyn et al. (2006) argued that these representations are not “depictive,” nor are they topographically organized on the cortex. The function of the Object-Properties Subsystem is to *recognize* objects in the Attention Window by comparing those incoming sensory stimuli against previously stored object properties.<sup>3</sup> Kosslyn (1994) stated that this subsystem corresponds to the ventral stream (“what”) pathway of the two-streams hypothesis (Goodale & Milner, 1992; Mishkin et al., 1983).

### ***Spatial-Properties Subsystem***

While Kosslyn's (1994) Object-Properties Subsystem (described earlier) is capable of recognizing objects, it discards information about location, size, and orientation. Individual points in the Attention Window are only important to the Object-Properties Subsystem insofar as

---

<sup>3</sup> Compare this with LIDA's sensory representations from Chapter 5.

they are arranged into particular shapes or patterns, not due to their locations in space. Kosslyn et al. (2006) argued that a separate subsystem—called the Spatial-Properties Subsystem—was needed to explicitly encode information about location, size, and orientation. This subsystem (which corresponds to neurons in the parietal lobes) encodes at least some of these spatial representations as topographically organized, depictive representations. Individual points are encoded within spatial representations relative to some designated origin within that reference frame (e.g., body-centric, object-centric, or scene-centric). This allows the system to represent the locations of objects, and parts of objects, within a topographically organized area. The objects at each of the locations encoded by the Spatial-Properties Subsystem could then be specified using the Object-Properties Subsystem. Kosslyn et al. (2006) referred to this style of hybrid representation as an *object map*<sup>4</sup>.

Kosslyn et al. (2006) hypothesized that the Spatial-Properties Subsystem encodes the locations of objects and object parts *across the entire visual field*. This is in contrast to the Object-Properties Subsystem that they argued was generally limited to encoding the portions of the visual field that were currently attended to by the Attention Window. Their rationale for supporting a wider encoder field for the Spatial-Properties Subsystem was to support attentional shifts to the locations of currently unattended objects within the visual field. The Spatial-

---

<sup>4</sup> Compare this with the cognitive “object” maps (Chapter 5) that were developed in the context of LIDA’s perceptual system.



Properties Subsystem corresponds to the dorsal stream pathway of an earlier (“what” vs “where”; Mishkin et al., 1983) formulation of the two-streams hypothesis<sup>5</sup>.

### ***Associative Memory***

The outputs (object maps and feature vectors) generated by the Object-Properties Subsystem and Spatial-Properties Subsystem are combined together to form encoded object and scene representations. These encoded (object- and spatial-property) representations are then compared against representations in short- and long-term associative memory. If an encoded representation matches a representation in associative memory, then the object it represents is said to be *identified*<sup>6</sup>. In the event that an object cannot be unambiguously identified, the best-matching representation is treated as a hypothesis (see Kosslyn et al., 2006, Chapter 5) that will guide later information gathering activities (see Information Shunting Subsystem).

Kosslyn (1994) stated that associative memory contains perceptual representations and more abstract conceptual representations (e.g., semantic knowledge). One final note is that Kosslyn et al. (2006) assumed that the output produced from associative memory would be a propositional description of an object or scene.

---

<sup>5</sup> Kosslyn (1994) explicitly identified the Object-Properties Subsystem and Spatial-Properties Subsystem as corresponding to the ventral and dorsal visual pathways; however, this characterization was later discarded by Kosslyn et al. (2006).

<sup>6</sup> Kosslyn et al. (2006) distinguished between “recognition,” which is performed by the Object-Properties Subsystem and “identification,” which is performed by associative memory. They stated that recognition merely indicates that a stimulus is familiar, whereas identification entails the ability to name an object and access other associated knowledge about that object (such as the categories to which the object belongs).

### ***Information Shunting Subsystem***

Whenever associative memory fails to clearly identify an object, the best-matching hypothesis is sent to an Information Shunting<sup>7</sup> Subsystem. This subsystem orchestrates a top-down search to confirm or refute an earlier hypothesis (e.g., from an ambiguous associative memory lookup). According to Kosslyn et al. (2006), the Information Shunting Subsystem serves two primary functions. First, it sends information to the Attention Shifting Subsystem. This information may then be used to adjust the focus of attention towards distinctive object parts or scene elements. Second, it sends information back to the Object-Properties Subsystem in order to *prime* various representations. This may, for example, facilitate the later encoding of various parts or characteristics.

### ***Attention Shifting Subsystem***

An Attention Shifting Subsystem updates the system's current focus of attention. This attentional shift can include changes to the location of the Attention Window, as well as orienting eye, head, and body movements. After attention is shifted, a new object or object parts may be encoded in the Visual Buffer. If those new stimuli match previously primed representations in the Object-Properties Subsystem, they may be recognized. Moreover, if the recognized objects or object parts clearly implicate associated representations (in associative memory), then those objects are said to be identified. Otherwise, the top-down search may continue based on new hypotheses, and additionally primed object parts and properties.

---

<sup>7</sup> Kosslyn (1994) referred to this as "information lookup." Kosslyn et al. (2006) renamed it to "information shunting."

## **Implementation: Mental Imagery in LIDA**

In this section, I develop a conceptual model of *multimodal* mental imagery and imagistic, epistemic processes in LIDA. Whenever possible, I compare this LIDA-based model with Kosslyn's (1994) proto-model of *visual* imagery and perception, which was described earlier in this chapter. At the end of this chapter, I also compare this implementation of mental imagery to those found in other cognitive architectures.

### ***Components***

Mental imagery utilizes the majority of LIDA's modules and processes. That said, several components stand out as being particularly significant. These are described in the subsections below.

**Sensory Scenes.** LIDA's Sensory Memory module may contain one or more *modality-specific* sensory scenes. A sensory scene is a data structure that preserves the low-level details of sensory stimuli, while simultaneously supporting the representation of more processed, low-level features. The representational content within LIDA's sensory scenes is typically integrated into *multimodal* representations elsewhere, for example, within LIDA's Current Situational Model.

A *visual* sensory scene (see Figure 25) could be represented as a layered, topographically organized data structure. Each layer within this data structure can represent features associated with active visual stimuli—real or virtual—as well as the locations in which those features occur within an agent's visual field. I refer to this representational format as a *feature map* (cf. LeCun et al., 1989; LeCun & Bengio, 1995).

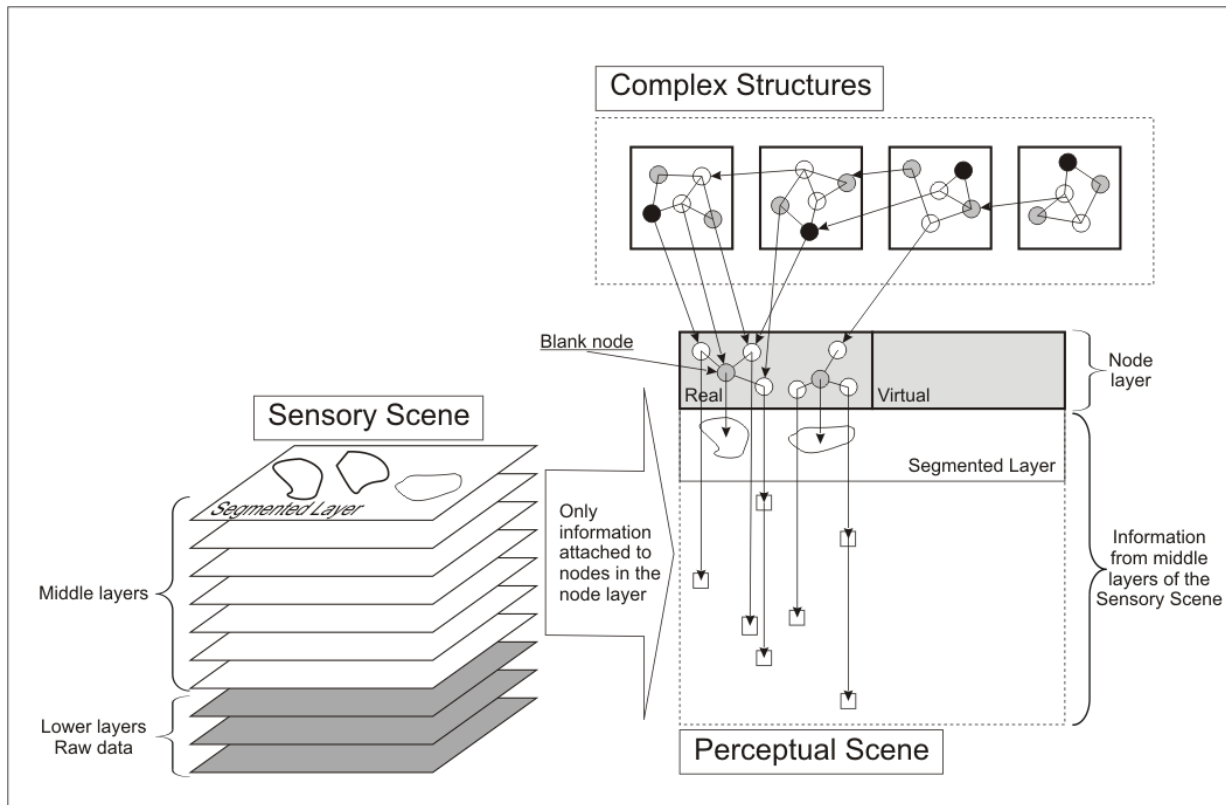


Figure 25. Visual sensory scene and Perceptual Scene. The visual sensory scene—a modality-specific, topographically organized data structure—exists in LIDA’s Sensory Memory module. Portions of this sensory scene are integrated into LIDA’s Perceptual Scene, specifically, those portions associated with current percepts. The Perceptual Scene is a multimodal data structure containing symbolic (nodes) and non-symbolic (depictive) representations. This figure originally appeared in McCall et al. (2010, fig. 3). (“Blank nodes” are coordinating amodal nodes associated with sensory content in the “segmented layer” of the visual sensory scene; see McCall, Snaider, et al., 2010 for more details.)

For software agents, the lowest layer of a visual sensory scene might be a “pixel” layer (McCall, Snaider, et al., 2010, p. 5) containing raw (unprocessed) sensory stimuli from an agent’s visual sensors (e.g., an RGB or depth camera). For natural agents (e.g., humans), it might contain a retinotopic map based on sensory projections within an agents’ eyes. Visual mental simulations (i.e., virtual sensory stimuli) could also be represented in this layer.

Higher layers in the visual sensory scene might contain feature maps derived from one or more of its lower layers—for example, color maps or motion fields (e.g., optic flow fields; Lee, 1980). These layers might, in turn, be used as the basis for even more processed layers higher in

the visual sensory scene. For example, an edge boundary feature map could be derived from the visual sensory scene's color and motion layers, which were, themselves, derived from its pixel layer.

LIDA's sensory scenes are purely *non-symbolic* data structures. As such, they are representationally different from Kosslyn's (1994) Visual Buffer, which contain symbolic (propositional) features overlaid on non-symbolic, spatial extents. Furthermore, sensory scenes are not directly accessible to LIDA's attentional processes; therefore, a direct analog to Kosslyn's (1994) Attention Window is impossible to implement within LIDA's Sensory Memory module. However, portions of these sensory scenes are accessible to attentional processes within LIDA's *Perceptual Scene*. Specifically, those portions with sensory representations attached to node structures in the Perceptual Scene's "node layer" (described below).

**Perceptual Scene.** LIDA's Perceptual Scene is a *multimodal* data structure within LIDA's Current Situational Model. It contains portions of LIDA's sensory scenes combined with node structures that characterize or identify that sensory content. These node structures are said to be part of the Perceptual Scene's *node layer* (McCall, Snaider, et al., 2010, p. 7). Modal nodes within the node layer are grounded in the representations within LIDA's sensory scenes. Note that sensory content corresponding to *multiple* modality-specific sensory scenes can be associated with a single node structure in the Perceptual Scene's node layer (see Figure 26). This is implemented using coordinating amodal nodes that support multimodal binding (see Chapter 5).

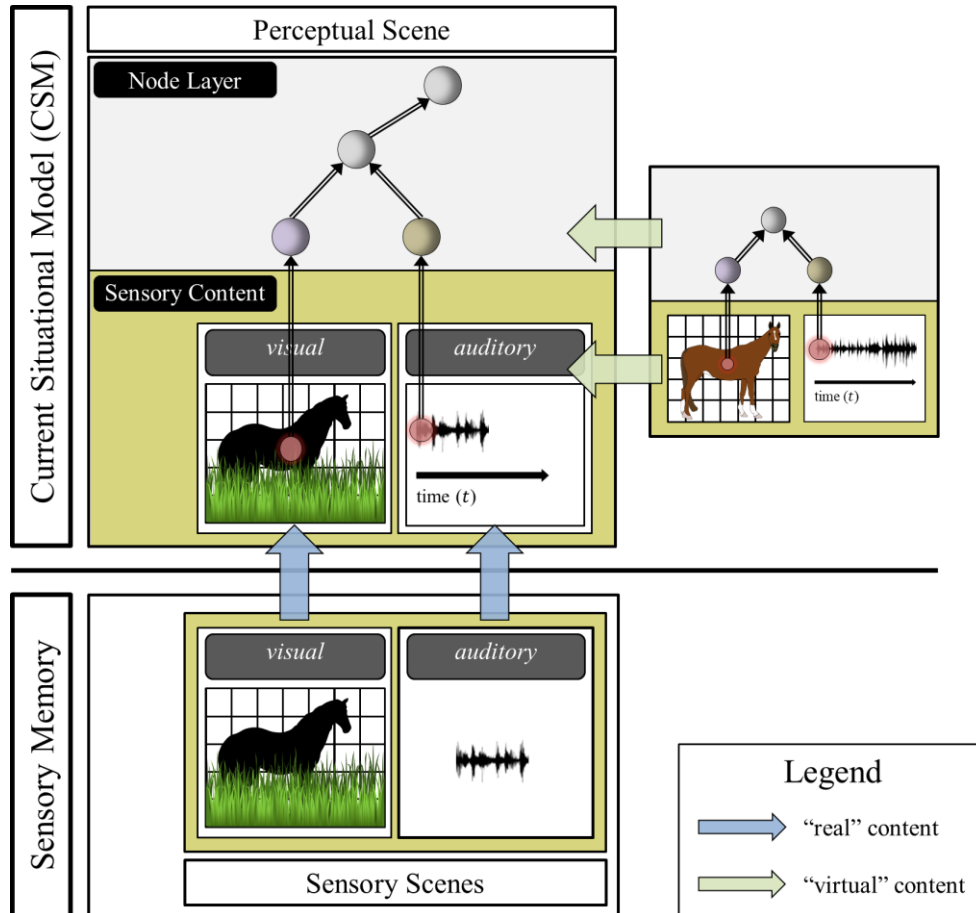


Figure 26. Real and virtual sensory content co-mingle in LIDA’s Perceptual Scene. Real sensory content, from Sensory Memory’s visual and auditory sensory scenes, is integrated into the sensory portions of LIDA’s Perceptual Scene. A percept, corresponding to the most likely interpretation of that sensory content (in this case a horse), is instantiated into the Current Situational Model. Virtual sensory content (i.e., a mental simulation) corresponding to this percept is generated and integrated into LIDA’s Perceptual Scene alongside real sensory content, which happens to be obscured and occluded. This virtual sensory content might serve to embellish the features of indistinct scene elements (among other things).

The Perceptual Scene can simultaneously contain “real” and “virtual” sensory content. Real sensory content originates from an agent’s external environment. Virtual sensory content is generated internally by “simulator” structure building codelets (see Chapter 5). During ordinary perception, virtual sensory content (top-down signals) augment real sensory content (bottom-up signals). These top-down signals help with the interpretation of that content, filling in missing details based on the most plausible explanations of that sensory stimuli. Both real and virtual

sensory content can be associated with perceptual/conceptual node structures in the Perceptual Scene's node layer.

Figure 26 illustrates the integration of bottom-up (real) and top-down (virtual) sensory signals. In the depicted situation, an agent encounters a partially occluded, equine creature of unknown variety at dusk. The “real” sensory content corresponding to this animal is partially occluded and contains indistinct features due to low-lighting conditions. Visual and auditory sensory stimuli resulting from this encounter are integrated into the agent's visual and auditory sensory scenes. This sensory content then activates perceptual representations in Perceptual Associative Memory (see Chapter 5, Multimodal Perception) resulting in the instantiation of a percept corresponding to a horse—the most plausible explanation<sup>8</sup> for that scene element. A simulator structure building codelet may then use this percept to construct a *spontaneous* (i.e., non-volitional) mental simulation of that horse, and integrate it into the sensory portions of the Perceptual Scene. This combined (bottom-up and top-down) signal could then be used to support active exploration or offline cognitive activities (e.g., generating action plans).

The Perceptual Scene is LIDA's closest analog to Kosslyn's (1994) Visual Buffer. Like the Visual Buffer, LIDA's Perceptual Scene is a hybrid representation. Its node layer contains symbolic or hybrid representations (i.e., node structures or cognitive maps) that characterize the non-symbolic representations in one or more sensory scenes. Furthermore, the Perceptual Scene can be operated on by LIDA's attentional processes (attention codelets), which can serve a

---

<sup>8</sup> Percept(s) instantiated from ambiguous sensory stimuli will typically depend on the base-level activations associated with node structures (i.e., recency and frequency effects), along with any residual current activations (priming) induced by recent situational contexts.

similar function to Kosslyn's (1994) Attentional Window. However, unlike the Visual Buffer, the Perceptual Scene is intended to serve as an integration point for sensory content from all sensory modalities. This introduces a degree of complexity—multimodal scene integration—that is not accounted for in Kosslyn's proto-model.

**Perceptual Associative Memory.** Perceptual Associative Memory (PAM) is LIDA's recognition memory. PAM is also the primary long-term memory module in which grounded, conceptual/perceptual representations are persisted. More elaborate forms of representational content (e.g., scenes, episodes, cognitive maps, and narratives) will typically depend on node structures from PAM. PAM may, in turn, depend on representations from other long-term memory modules. For example, cognitive "object" maps, which are analogous to Kosslyn's object maps, are formed through the association of PAM nodes with object-centered spatial extents from Spatial Memory (see Chapter 5, Cognitive "Object" Maps).

Perceptual Associative Memory (PAM) combines aspects of Kosslyn's Object-Properties Subsystem and Associative Memory. PAM is like Kosslyn's Object-Properties Subsystem in the following ways: First, the activation of its representations can be based on object properties (e.g., color, texture, and shapes). Furthermore, PAM's representations (e.g., node structures) are not topographically organized, and they do not directly "depict" their referents. However, through their associations with other forms of long-term memories (e.g., representations in Spatial Memory) they can form spatially arranged, depictive, hybrid representations (e.g., cognitive "object" maps). Finally, PAM's grounded node structures could be described as population codes: their activations are based on a set of nodes and a set of sensory representations that characterize sensory stimuli.



There is one notable difference between Kosslyn's Object-Properties Subsystem and PAM: attention modulates their operations differently. According to Kosslyn, his Object-Properties Subsystem is constrained to the recognition of objects within an Attention Window. In contrast, the bottom-up activation of PAM is unaffected by LIDA's attentional processes. However, attentional processes (attention codelets) that function like Kosslyn's Attention Window can be applied *after* the activation of PAM nodes (from Sensory Memory), that is, once percepts are instantiated in the Current Situational Model. It is very likely that this difference could lead to testable predictions. For example, even if preconscious percepts outside of the focus of attention are never brought to consciousness, they may alter Workspace dynamics, leading to priming effects and other secondary indicators.

Finally, note that PAM models some, but not all, elements of Kosslyn's Associative Memory. In particular, Kosslyn's Associative Memory contains semantic knowledge in addition to perceptual knowledge. In LIDA, semantic knowledge is encoded in a separate Declarative Memory module. Also, Kosslyn assumed that the output of Associative Memory was symbolic propositions. PAM's output, on the other hand, typically consists of hybrid (symbolic/non-symbolic) representations that can be said to characterize scene elements, but not propositionally.

**Action Selection.** LIDA's Action Selection module plays a pivotal role in supporting many forms of mental imagery. For example, we saw in Chapter 6 that "motor imagery" (Grush, 2004; Jeannerod, 1994, 1995, 2001; Pfurtscheller & Neuper, 1997; Sharma & Baron, 2013) is based on the selection of behaviors for internal execution (i.e., "covert actions"; Jeannerod, 2001, p. S103). Furthermore, there is a sizeable body of research supporting the hypothesis that image

transformation (e.g., mental rotation) relies on internal-action-based motor processes (Ganis et al., 2000; Georgopoulos et al., 1989; Kosslyn et al., 2001; Wexler et al., 1998).

More generally, Action Selection orchestrates all *action-based mental imagery*, which includes not only motor imagery, but all forms of voluntary or intentional mental imagery (see Albright, 2012; Brewin et al., 2010; Kreiman et al., 2000). This may even include some forms of mental imagery that are construed as spontaneous or involuntary. In particular, mental imagery that relies on automatized and consciously mediated modes of action selection may not be perceived as intentional acts by an agent. Consequently, Action Selection may frequently participate in every fundamental operation of mental imagery—generation, transformation, inspection, and maintenance (Kosslyn, 1994; Kosslyn et al., 2006).

**“Simulator” Structure Building Codelets.** In LIDA, all *preconscious* mental simulations are generated by simulator structure building codelets (SBCs) (see Chapter 6, Simulator Structure Building Codelets). This applies to mental simulations generated spontaneously (e.g., following the recall of long-term memories), as well as volitional or intentional mental simulations based on covert (internal) actions. However, *never conscious* mental simulations, such as those produced by “forward models” (D. M. Wolpert et al., 1995) in LIDA’s Sensory Motor System, are *not* based on simulator SBCs.<sup>9</sup> SBCs only operate on representations in the (preconscious) Workspace.

---

<sup>9</sup> Dong et al. (2015) implemented a simple forward model in LIDA’s Sensory Memory System using a Kalman filter-based approach.

Simulator structure building codelets (SBCs) generate modal simulations from grounded concept representations in LIDA's Current Situational Model (CSM). That is, before a simulator SBC can create a modal simulation, some cognitive process must introduce a grounded node structure into the CSM. Cognitive processes capable of this include perception (see Chapter 5, Multimodal Perception), the cueing of long-term memory modules from the CSM (see Chapter 4, The LIDA Cognitive Cycle), the internal execution of behaviors (see Chapter 6), and other structure building codelets.

$\beta$ -VAEs (Higgins et al., 2017) provide one computational implementation for simulator SBCs (described in Chapter 5) However, this is by no means the only implementation. Regardless of the implementation used, the act of mental simulation should be based on the re-activation of LIDA's sensory representations via coordinating node structures in Perceptual Associative Memory (PAM). Moreover, this re-activation should be performed in a top-down fashion—for example, from PAM nodes with higher conceptual depth, to those with lower conceptual depth, eventually re-activating sensory representations (see Chapter 5, Simulator Structure Building Codelets).

**Attention Codelets.** Attention codelets are LIDA's attentional processes. They scan preconscious representations within the Current Situational Model and bring any content of interest to them to LIDA's Global Workspace (via a coalition forming process). If such content is included in a winning coalition, it is consciously broadcast to all of LIDA's modules and processes (see Chapter 4, The LIDA Cognitive Cycle).

A multimodal version of Kosslyn's Attention Window could be implemented by one or more attention codelets that are interested in aspects of LIDA's Perceptual Scene. For example,

specific regions within the visual sensory scene<sup>10</sup> could be attending to by one or more attention codelets. However, a fundamental question is, what representations and/or cognitive processes specify or modulate these focal parameters (e.g., the size and shape of the area of interest in the visual sensory scene)? Furthermore, some of these parameters appear to be amenable to volitional control (i.e., based on intentional, internal actions). These topics will be addressed later in the chapter (e.g., with respect to LIDA’s implementation of “image inspection”).

### ***Fundamental Operations of Mental Imagery***

This section describes LIDA-based implementations of image generation, image transformation, image inspection, and image maintenance—Kosslyn’s (1994) four fundamental mental imagery operations. These implementations are based on the assumption that the linear timing effects observed during many imagery-based chronometric studies is a direct consequence of *multi-cyclic* offline cognitive processes (see Chapter 4, The LIDA Cognitive Cycle). Moreover, I assume that these cognitive processes are dependent on internally executed behaviors.

**Image Generation.** Image generation refers to the “spontaneous” or “intentional” formation of mental simulations. Spontaneous image generation is characterized as being “involuntary” (Brewin et al., 2010) or “implicit” (Albright, 2012). While intentional image generation is characterized as “voluntary” (Brewin et al., 2010), “volitional” (Kreiman et al., 2000) or “explicit” (Albright, 2012). In both cases, the corresponding (preconscious) mental simulations are generated by simulator structure building codelets (see Chapter 5, Simulator Structure

---

<sup>10</sup> Note that the representational content in sensory scenes is only accessible to processes (e.g., codelets) operating on the CSM if that content that been associated with a node structure (grounded concept representation) in the Perceptual Scene’s node layer (see Figure 25).

Building Codelets) and integrated into LIDA's Perceptual Scene. Both spontaneous and voluntary forms of image generation are depicted in Figure 27.

Spontaneous mental simulations can occur in LIDA whenever a percept or cued memory enters the Current Situational Model (CSM) and is subsequently acted upon by a simulator SBC. Content generated in the CSM from the *consciously mediated* or *automatized* selection of internally executed behaviors may also support spontaneous mental simulations. In these cases, the results of behaviors<sup>11</sup> selected for internal execution function like cued long-term memories.

In contrast, intentional mental simulation is *always* initiated by an internally executed, covert behavior. Furthermore, these behaviors are selected, at least in part, due to an explicit intention to engage in mental simulation. This suggests that the mode of action selection that supports intentional mental simulation is volitional (i.e., goal-directed), but not necessarily deliberative<sup>12</sup>. In both cases, simulator structure building codelets are directly responsible for orchestrating the creation of these preconscious mental simulations.

Single-part mental simulations can be generated in a single cognitive cycle. This form of image generation can occur completely preconsciously (e.g., from percepts or cued memories), or it can be generated following the execution of covert behaviors (e.g., motor simulations).

---

<sup>11</sup> Recall from Chapters 4 and 6 that behaviors are instantiated schemes. Schemes are data structures that are learned into Procedural Memory that contain three primary components: a context, an action, and a result. Action-based mental simulation relies extensively on schemes' (instantiated) results.

<sup>12</sup> The LIDA literature currently does not make a distinction between volitional action selection and deliberation (Franklin et al., 2016, sec. 6.2). However, in many cases of intentional mental imagery, individuals appear to have explicit intentions to perform mental imagery, but the corresponding "options" to act do not initiate a deliberative context. Furthermore, the selection of such actions appears to occur in a single cognitive cycle. This suggests that these "intentional" actions should either be modeled in LIDA as consciously mediated actions, or that LIDA's volitional mode of action selection needs to be updated to include the non-deliberative but intentional selection of actions.

Multi-part mental simulations, on the other hand, likely *require* one or more supporting internal actions. In these cases, the incorporation of additional parts and properties occurs largely through a process of *sequential elaboration*, in which parts and properties are added one at a time (see Kosslyn et al., 1983, 1988). During sequential elaboration, the cognitive processes responsible for inspection, transformation, and generation collaborate to construct parts and properties on demand.

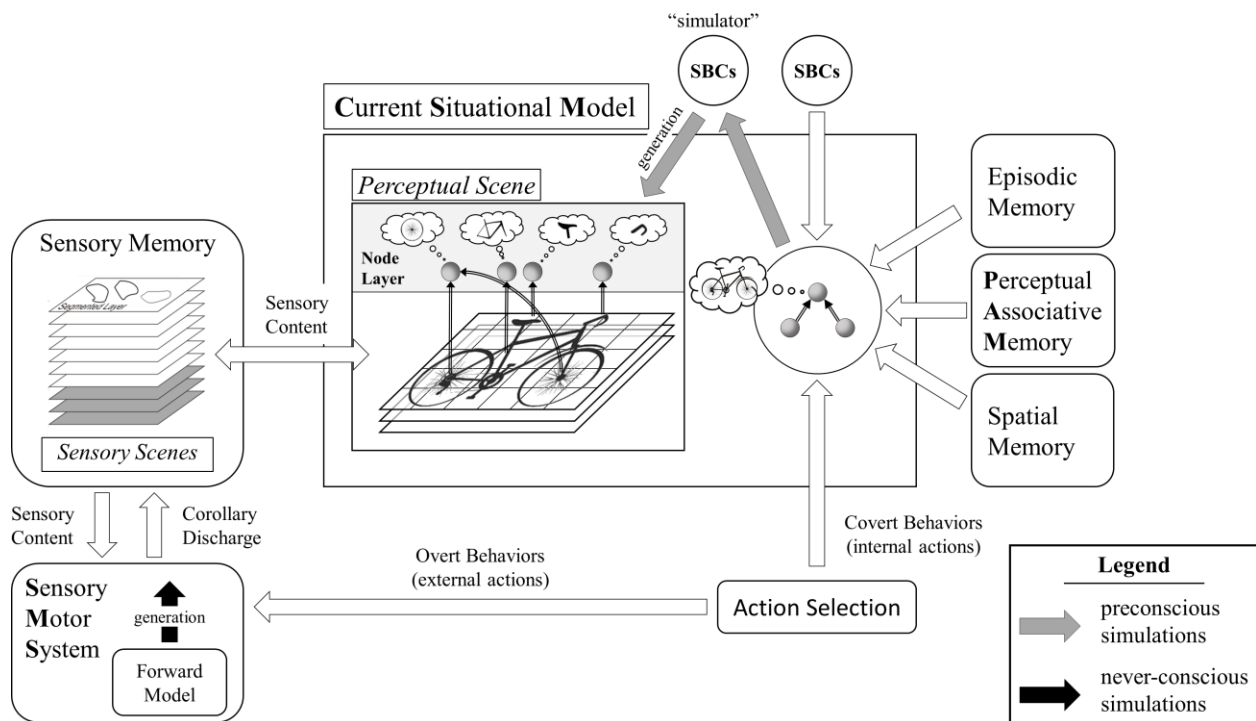


Figure 27. Mental imagery and the Perceptual Scene. Spontaneous forms of mental imagery can occur during bottom-up perception, and following the cueing of long-term memory. Intentional forms of mental imagery are always action-based, using covert behaviors (with internal actions) to initiate changes within LIDA’s Current Situational Model. All forms of preconscious mental simulation are supported by “simulator” structure building codelets (SBCs) that generate mental simulations and integrate them within LIDA’s Perceptual Scene. Never-conscious mental simulations can occur in the Sensory Motor System (SMS) as a result of external (overt) action execution. The Perceptual Scene can contain both “real” and “virtual” sensory and perceptual representations.

**Image Transformation.** The mechanisms proposed here for implementing image transformations are largely the same as those used for motor imagery (described in Chapter 6); that is, image transformations are modeled as *action-based* imagistic processes. Specifically, image transformations (e.g., mental rotations) typically involve the serial execution of consciously mediated or volitionally selected internal actions. These covert actions can then initiate the generative activities of unconscious processes (e.g., simulator structure building codelets).

The hypothesis that image transformations are action-based is supported by neurophysiological research that shows the engagement of motor processes during mental transformations (Cohen et al., 1996; Georgopoulos et al., 1989; Kosslyn et al., 2001; Wexler et al., 1998; Zacks, 2008). In particular, the premotor cortex, which is believed to be involved in the high-level selection and preparation of voluntary movements<sup>13</sup> (Passingham, 1988; Scott & Kalaska, 2021) is typically activated during image transformations, such as mental rotations (e.g., see Cohen et al., 1996; Kosslyn et al., 2001). The premotor cortex is functionally analogous to LIDA’s Action Selection and Procedural Memory modules; therefore, modeling image transformations as being action-based is consistent with current experimental research.

If image transformations are, indeed, action-based, then one might expect that their development and use would be similar to that of motor skills—and this seems to be the case. For example, numerous studies have shown improvements in mental imagery abilities with age

---

<sup>13</sup> The premotor cortex is also believed to support the *understanding* of others’ actions (Buccino et al., 2013; Gallese et al., 1996; Rizzolatti et al., 1996). This suggests that “mirror mechanisms” (Iacoboni et al., 2005; Rizzolatti & Craighero, 2004; Rizzolatti & Sinigaglia, 2016) could also be implemented in LIDA using internally executed behaviors.

(Caeyenberghs et al., 2009; Kosslyn et al., 1990; Smits-Engelsman & Wilson, 2013; Souto et al., 2020) and training (Spruijt et al., 2015). And motor imagery training appears to improve motor performance (see Driskell et al., 1994 for a meta-analysis).

In general, the view advocated for here is that mental imagery represents the internalization of an individual's interactions with the world (Piaget & Inhelder, 1971). For example, the conscious awareness of the consequences of one's actions allows those consequences to be internalized as schemes in Procedural Memory. Through repeated use, these learned behaviors can be internally (covertly) executed—instead of their overt counterparts—in service of perception and offline cognition.

Frick et al.'s (2009) findings support this gradual transition from overt to covert activity. They found that young children (5-year-olds) relied on concurrent, overtly executed movements when trying to simulate the consequences of imagined events (e.g., changes in the orientation and position of imaginary water when tilting an empty glass). However, older children (e.g., 9-year-olds) were less dependent on these overt behaviors. Frick et al. (2009) suggested that one explanation for these results is that older individuals may be able to mentally simulate their hand movements, whereas children required overt hand movements to facilitate their mental simulations. As individuals become more experienced, they became more capable of using covertly executed behaviors in place of overt behaviors.

The internalization of overt behaviors can be observed in players of strategy board games, like chess or Go. For example, beginning chess players will often physically move pieces on the board to visualize the consequences of their moves and assess board positions. However, as they gain experience, and begin playing in more constrained settings where those overt



behaviors are prohibited, they often develop the ability to perform these moves mentally (i.e., covertly).

Recall that LIDA has four modes of action selection: volitional, consciously mediated, automatized, and alarms (see Chapter 4). Throughout this work, I have generally assumed that volitional and consciously mediated action selection is compatible with the internal execution of behaviors. For example, intentional and spontaneous image generation (described earlier in this chapter) seems to require these modes of action selection. In contrast, I argued in Chapter 6 that alarms are *incompatible* with the internal execution of behaviors, due to the time-sensitive nature of the situations in which they arise. However, up to this point, I have not made any explicit declarations about whether *automatized* (autonomous; Fitts & Posner, 1967) action selection could occur with mental imagery.

LIDA's automatized action selection is characterized as being largely unconscious, where one selected behavior seems to directly call the next (in a stream of behaviors) without the need for intervening conscious content. This mode of action selection is typically only available after an individual has mastered a skill, and, even then, only in predictable situations that do not require conscious intervention.

The model of mental imagery developed in this manuscript predicts that internal behaviors *should* be compatible with automatized action selection. Unlike alarms, there is no criteria to exclude automatized behaviors from being internally executed. Furthermore, this capability may be highly advantageous—e.g., expediting an agent's offline cognitive processes by reducing their dependence on conscious content during action selection.

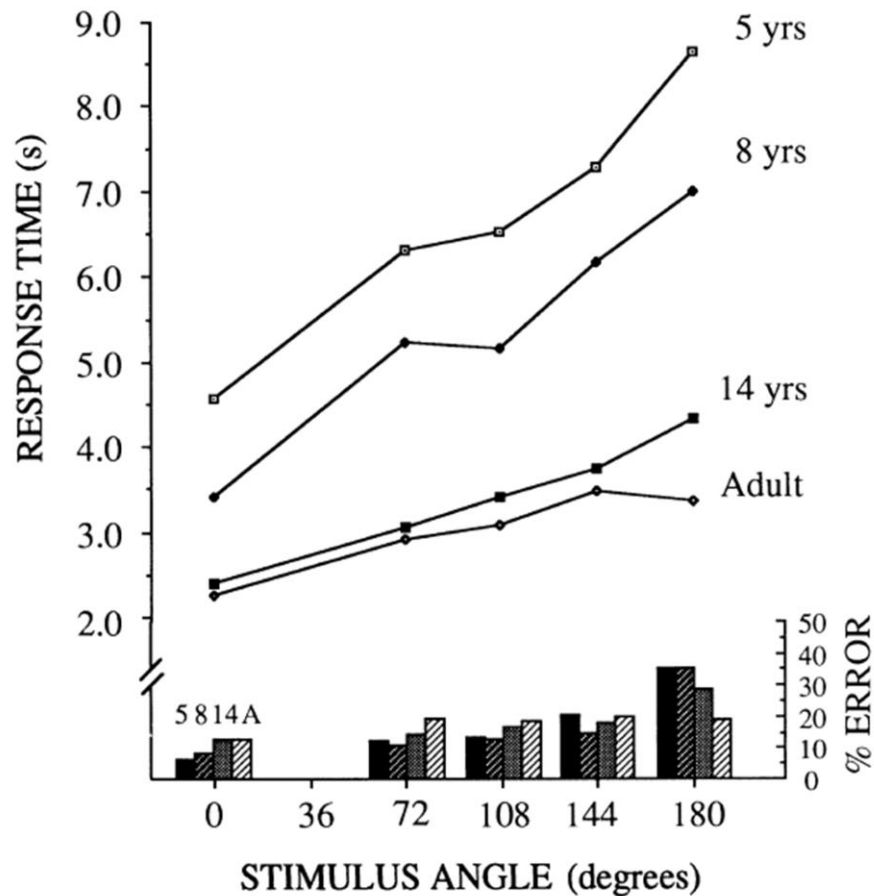


Figure 28. Response times in mental rotation across age groups. All age groups exhibited response times that had a linear relationship with respect to stimulus angle; however, the slopes associated with the corresponding regression lines decreased noticeably with increased age. Figure reprinted from Kosslyn et al. (1990, fig. 4) with permission.

Some support for this “internal automatization hypothesis” comes from experiments by Kosslyn et al. (1990), who performed a modified version of Shepard and Metzler’s (1971) mental rotation experiment, including a chronometric study of the mental rotation abilities of 5-year-olds, 8-year-olds, 14-year-olds, and adults. While their results showed a predictable, linear relationship between subjects’ response times and the angle of stimulus presentation for all age groups, the *slopes* of their corresponding regression lines decreased considerably with age (see Figure 28). It is unclear what form of learning or development accounts for these improvements

in efficiency; however, one possibility is that there are underlying “behavior streams” (see Chapter 6) supporting mental rotation that are progressing from consciously mediated to automatized forms of action selection (cf. cognitive, associative, and autonomous stages of skill acquisition; Fitts & Posner, 1967).

Madl et al. (2011) estimated that the time required for a single cognitive cycle is approximately 260-390 milliseconds. The linear regression line for adult subjects in this experiment showed a slope of 245 milliseconds per 36 degrees of stimulus rotation. Unless subjects were performing covert mental rotations in increments of 36 degrees or more, these results suggest that parts of these operations may be occurring unconsciously (i.e., without conscious mediation). Targeted experiments are needed to explore this possibility further.

One final implementation detail that I will consider here relates to the manner in which behaviors with composite actions—i.e., behaviors whose actions are implemented using behavior streams—are mentally simulated. Recall that the internal execution of behaviors with *primitive* actions involves the immediate simulation of those behaviors’ results. That is, following the selection of such a behavior for internal execution, a structure building codelet will update LIDA’s Current Situational Model with that behavior’s expected result. A simulator structure building codelet might then generate a mental simulation of that result and integrate it into LIDA’s Perceptual Scene.

In contrast, the internal execution of behaviors with *composite* actions could be modeled in one of two ways: (1) their predicted results could be simulated directly (bypassing their underlying behavior streams), or (2) the component behaviors in their underlying behavior streams could be simulated iteratively (and potentially recursively). The chronometric studies

considered throughout this chapter (e.g., Borst et al., 2006; Cooper, 1975, 1976; Finke & Pinker, 1982; Kosslyn et al., 1978; Metzler & Shepard, 1974; Pinker et al., 1984; Shepard & Metzler, 1971) suggest that the latter is occurring—reaction time increases with angle of rotation, scanning distance, object complexity (i.e., the number of parts in a multi-part image), etc. These timing effects would not be predicted if the automatized behavior’s expected result were directly simulated.

Furthermore, this manner of internal execution is predicted by the implementation developed in Chapter 6. In particular, whenever a behavior with a composite action is selected, Action Selection *immediately* chooses a component behavior from its underlying behavior stream. And it is this component behavior’s result that would be mentally simulated into LIDA’s Current Situational Model, not the high-level behavior’s result.

**Image Inspection.** Image inspection entails the exploration and examination of the sensory content in LIDA’s Perceptual Scene. It relies on both imagistic operations and perceptual processes. While image inspection often coincides with image transformations (e.g., during mental scanning; Kosslyn et al., 1978) and image generation (e.g., during the sequential elaboration of multi-part images; Kosslyn, 1994, Chapter 9), it is functionally and neurologically distinct (Kosslyn et al., 2004).

In this section, I propose an implementation of image inspection in LIDA based on *action-mediated* attention processes—i.e., attentional processes modulated by the *intentional* execution of *orienting behaviors*. Posner defined “orienting” as the alignment of attention to a source of sensory stimuli or internal representations (Posner, 1980, p. 4). Orienting is distinct from the cognitive processes responsible for the conscious awareness of those stimuli and

representations, which Posner referred to as “detecting” (Posner, 1980, p. 4). For example, someone may move their eyes towards a stimulus, and yet be unaware of it. Nevertheless, orienting towards something often precedes the conscious awareness of that thing.

In LIDA, “detecting” is implemented by attention codelets and the Global Workspace. In contrast, “orienting” involves the execution of behaviors that modulate the operations and concerns of attention codelets. This may involve spawning new attention codelets or directing their interests towards specific portions of LIDA’s Perceptual Scene (or types of content within it).

Orienting behaviors can either be overt or covert. Eye and head movements are examples of overt orienting behaviors. However, attention can also be oriented through purely covert means, using internal mechanisms alone. For example, an individual may be looking directly at an object, and yet attending to something in their peripheral field of vision (e.g., see Posner et al., 1978).<sup>14</sup>

Recall that Kosslyn’s (1994) proto-model includes an Attention Window that is used to delineate an agent’s visual locus of attention. Kosslyn described the Attention Window as a contiguous set of points that could be overlaid on portions of the Visual Buffer. As such, the Attention Window enables the separation of the focus of visual attention from foveal sensory content, which would presumably be in the center of the Visual Buffer during normal perception.

---

<sup>14</sup> Though non-visual modalities have received far less research, orienting mechanisms have been demonstrated for other sensory modalities, including auditory (Mondor & Zatorre, 1995; Spence & Driver, 1994) and tactile (Spence & Gallace, 2007) modalities.

Moreover, Kosslyn's Attention Shifting Subsystem provides the means for executing overt and covert orienting behaviors that adjust the position and shape of the Attention Window.

Hoffman and Subramaniam (1995) conducted a series of experiments in which they investigated the relationship between saccadic eye movements and the covert orienting of visuospatial attention. They interpreted the results of their study to suggest that overt, orienting behaviors (e.g., saccadic eye movements) *depend* on preliminary, covert, orienting behaviors. This view is broadly consistent with Rizzolatti's "premotor theory of attention" (Rizzolatti et al., 1987), which posits that overt and covert orienting behaviors are controlled by common underlying mechanisms. In both cases, a "motor program" is generated; however, physical movements (e.g., eye movements) are blocked by the nervous system during purely covert executions. The premotor theory of attention is supported by numerous neuroimaging-based (e.g., fMRI) studies that show substantial overlap in neural activations when subjects performed covert and overt orienting behaviors (Beauchamp et al., 2001; de Haan et al., 2008; Nobre et al., 2000). This theory also accords with Jeannerod's (2001) theory of motor cognition, which was introduced in Chapter 2.

Finally, orienting behaviors can be either reflexive or volitional. For example, eye movement can be driven unintentionally (reflexively) based on environmental stimuli (e.g., unexpected moment in the peripheral visual field), or intentionally (volitionally) based on a premeditated oculomotor trajectory. Reflexive forms of orienting are often referred to as "exogenous," while volitional forms of orienting are referred to as "endogenous" (Carrasco, 2011). These two modes of attention seem to follow different time courses (Busse et al., 2008) and appear to be based on different cognitive processes and neural systems (Chica et al., 2013).

Furthermore, reflexive (exogenous) forms of attentional shifts are believed to be unavailable during mental imagery (Kosslyn, 1994, p. 102); therefore, the focus of this section will be on *endogenous* attentional shifts.

There are several options for implementing endogenous (i.e., intentional), covert orienting behaviors in LIDA. All of these options depend on internally executed behaviors to initiate covert attentional shifts, and the associated changes within the preconscious Workspace modulate LIDA's attentional processes (i.e., attention codelets). Each option considered proposes a different mechanism for effecting the necessary attentional biases—i.e., how attention is focused on specific content within LIDA's Perceptual Scene.

The first option involves spawning or activating attention codelets that are focused on specific representations in LIDA's Current Situational Model; for example, they may be concerned with visual sensory content contained in a bounded region of LIDA's Perceptual Scene (cf. Kosslyn's Attention Window). These attention codelets are created based on internally executed orienting behaviors. This mechanism is similar to LIDA's use of "expectation codelets" (D'Mello et al., 2006) to bias attention towards anticipated action consequences.

A second option involves the creation of transient, preconscious representations (e.g., "attention-directing" nodes) that modulate LIDA's attentional processes. For example, structure building codelets could be spawned (or activated)—via internally executed behaviors—that augment specific preconscious representations with associated (linked-to) attention-directing representations. Corresponding attention codelets could then scan the Workspace, looking for structures containing that "decorated" content. Coalitions containing these structures could then be brought to the Global Workspace to compete for inclusion in a global broadcast.

A third, and final, option requires the introduction of a persistent data structure within the Perceptual Scene that specifies an agent's current attentional foci. This data structure would function like a multimodal version of Kosslyn's Attention Window, continually directing the activities of a set of attention codelets towards content demarcated by those attentional foci. This data structure could be viewed as an enhancement to LIDA's Perceptual Scene that allows an agent's attentional concerns to be explicitly represented with respect to the contents of the Perceptual Scene. Internally executed orienting behaviors could then be used to adjust the parameters (e.g., the location and extent) of those attentional foci.

At this point, it is unclear which of these options (if any) is preferable. Option three (i.e., a multimodal "attentional window") could most easily support the incremental, attentional shifts observed in many experiments based on visual modalities (Cave & Kosslyn, 1989; Eriksen & Murphy, 1987; Eriksen & St. James, 1986; Kosslyn, 1994, p. 94). However, Posner et al. also argued that orienting is an "active process" (Posner et al., 1984) that requires continual maintenance, rather than a passive filter that can be "set in place and left" (Posner, 1980, p. 8). In contrast, options one and two are "active processes" that require continual maintenance, but it is less clear how they could support incremental attentional shifts—e.g., what internal state is being maintained and shifted? As such, it is possible that some combination of these options should be used. What is clear is that each of these options makes different predictions that require additional scrutiny, however, that exercise is beyond the scope of this manuscript.

**Image Maintenance.** All of LIDA's representations and codelets are subjected to activation decay (see Franklin, Strain, et al., 2013, sec. 4.7), and, in the absence of sufficient reinforcement, they are eventually purged from an agent's short- or long-term memory modules. Image



maintenance is, therefore, required to ensure that the sensory and perceptual representations associated with mental simulations—in LIDA’s Perceptual Scene—remains available to imagistic and perceptual processes.

Kosslyn hypothesized that the maintenance of (visual) mental images is based on the re-activations of “compressed image representations” and attentional processes (Kosslyn, 1994, p. 325). He also argued that image maintenance *does not* depend on image generation, at least not the processes “that are used to integrate parts into an image during image generation” (Kosslyn, 1994, p. 321). This position was based, in part, on the findings of neuroimaging studies that show a lack of activation in brain areas associated with image generation (e.g., Uhl et al., 1990).

Additional arguments against full-blown image generation could be made based on the variability inherent in mental simulation. Recall that mental simulation is a dynamic process (see Chapter 5), where each generative event typically produces different sensory content. With respect to the current computational implementation, this variability can occur as the result of several factors. The most direct cause is the use of a probabilistic *sampler* to implement the simulator structure building codelet (see Chapter 5, Simulator Structure Building Codelets). Depending on the magnitude of the standard deviations<sup>15</sup> associated with a sensory representation (i.e., modal probability distributions), even single-part reconstructions can vary considerably (see Figure 34). Multi-part reconstructions would compound this issue. Therefore, if image generation were involved in image maintenance, there would likely be perceptible alterations in an underlying mental simulation following each generative event.

---

<sup>15</sup> Recall from Chapter 5 that  $\beta$ -VAEs produce sensory representations that are composed from a *vector* of means ( $\vec{\mu}$ ) and a *vector* of standard deviations ( $\vec{\sigma}$ ).

A more significant source of variability can occur due to changes in the activations within PAM. Recall that the simulation of grounded concept representations (see Chapter 5) requires top-down iteration—from nodes with higher to lower conceptual depth—until modal nodes containing sensory representations are activated. Depending on the activation dynamics within PAM, the specific tokens (i.e., category members) that are selected for simulation may change along with their properties (e.g., colors and textures).

Given the above, I propose a simple implementation of image maintenance based on the direct reinforcement of preconscious Workspace activations. In particular, conceptual/perceptual nodes and their associated mental simulations in the Perceptual Scene can be reinforced if that content is included in a global broadcast. The global broadcast is received by all of LIDA's modules, including the preconscious Workspace; therefore, the Workspace could use that content to increase the activations associated with its preconscious representations (including mental simulations). This is supported by the execution of internal “orienting” behaviors that bias an agent's attention towards specific representations and their mental simulations—thus, increasing the likelihood that they will come to consciousness and be reinforced.

As a direction of future work, the relationship between image maintenance and working memory (Baddeley, 1992; Baddeley & Hitch, 1974, 1994) should be explored. In particular, modality-specific forms of working memory (e.g., visual working memory; Eng et al., 2005; Fougne et al., 2012; Luck & Vogel, 1997; Vogel et al., 2001) and their relationship to mental imagery should be examined more deeply.

## **A LIDA-Based Agent**

A simulation-based LIDA agent was designed to be a “participant” in a perceptual experiment inspired by Kosslyn et al. (1990). While this agent was only partially implemented in software, its design and preliminary results provide a useful demonstration of the concepts developed throughout this manuscript.

The intent of this agent and its environment was to show how—through the overt execution of *epistemic behaviors*<sup>16</sup>—the consequences of an agent’s actions could be internalized, enabling the generation of experience-based predictions (e.g., action-based mental simulations) that support perception. It was also intended to show how the various perceptual, procedural, and imagistic processes detailed throughout this manuscript could be combined to implement a complete autonomous software agent.

### ***Environment***

Kosslyn et al. (1990) described a mental rotation experiment (referred to as “task 4”) in which participants were shown two panels. Each panel contained asymmetrical, connected shapes formed from randomly selecting five cells in a 4 x 5 regular grid (i.e., pentominoes; see Figure 29). The shape in the left panel was always displayed upright and served as a reference image. The shape in the right panel was always rotated from its upright position (in multiples of 36°).

---

<sup>16</sup> Kirsh and Maglio (1994) defined *epistemic actions* as actions that are primarily intended to discover information or simplify problem-solving. Temporarily moving a chess piece to a new position on a chess board—to visualize the consequences of a candidate move—is an example of this. In contrast, they defined *pragmatic actions* as actions that are intended to directly bring an agent closer to its goals.

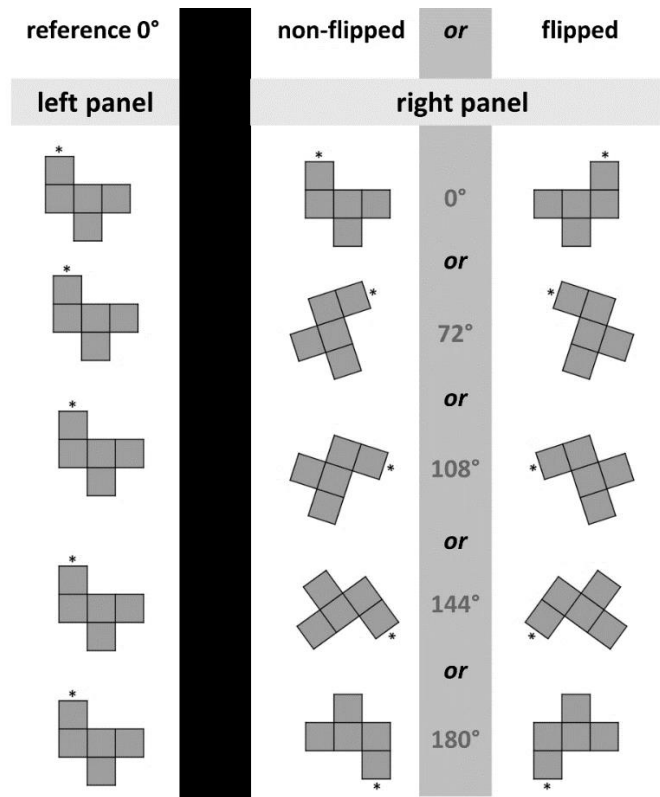


Figure 29. Materials and procedures used in mental rotation experiment. Kosslyn et al. (1990) referred to this experiment as Task 4.

On half of the trials, participants were shown a shape in the right panel that was a rotated copy of the shape in the left panel. On the other half of the trials, they were shown a rotated, *mirror image* of the shape. Asterisks were placed at the corresponding tops of both stimuli, to minimize the difficulty of discovering the relative orientations of the two shapes.

Figure 29 shows an example, upright shape (such as would appear in the left panel) along with several combinations of rotated or reflected *and* rotated shapes (such as those that would appear in the right panel). Participants were asked to decide whether the shapes shown in both panels were identical, regardless of their orientation. However, they were instructed that only two-dimensional rotations were allowed, “as would occur if a pattern were on a piece of paper on a table top and one could not lift it off the table” (Kosslyn et al., 1990, p. 1007).

A software environment (see Figure 30) was developed based on Kosslyn et al.’s (1990) experiment (“task 4”). This environment generates two 128 × 128 pixel, black-and-white images (one for each panel), and an agent’s task is to determine if the shapes in the left and right panel are the same. However, once an agent has made its guess, the environment provides *no feedback* to the agent as to the correctness of its shape classifications. It merely cycles to the next pair of shapes. Moreover, a few modifications were made to Kosslyn et al.’s (1990) experimental procedure to make these perceptual task more interesting and challenging for a software agent.

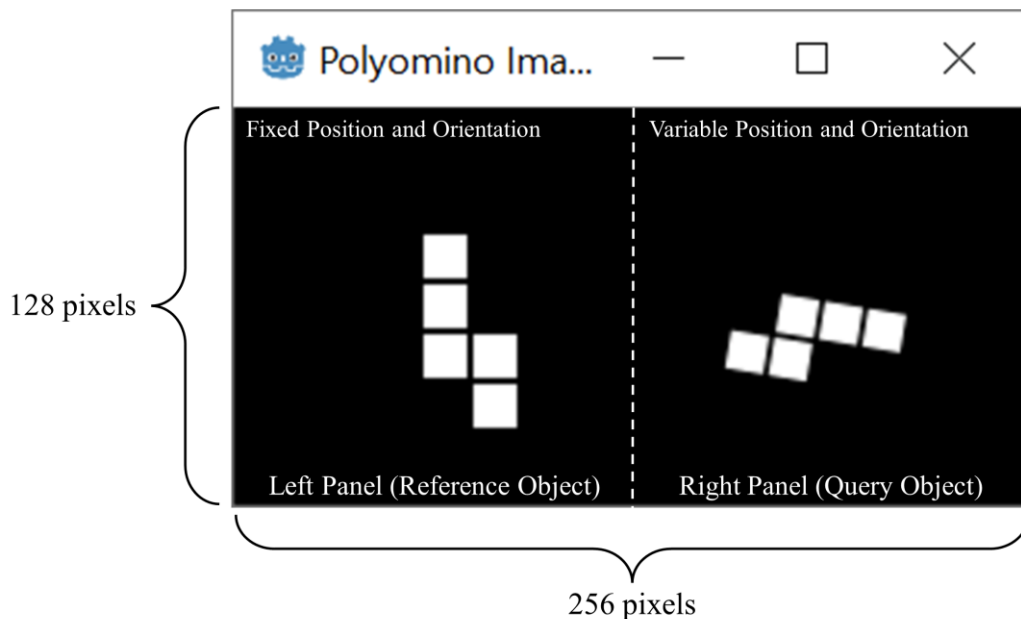


Figure 30. Screenshot of the mental rotation experiment’s environment.

The first change was to support 29 *polyominoes*—rather than the 12 *pentominoes*<sup>17</sup> supported by Kosslyn et al. (1990). These were composed by selecting five *or fewer* contiguous

---

<sup>17</sup> Kosslyn et al. stated that they used *nine* different stimulus shapes (1990, p. 1007), but it is unclear how they arrived at this number. There are 18 pentominoes, in total, but only 12 of them are asymmetrical under reflection—i.e., their reflections produce distinct shapes. This leads to six stimulus shapes and six mirror images of those shapes. Therefore, I report their number of stimulus shapes as 12 here (rather than nine), but this does not necessarily mean Kosslyn et al.’s (1990) reporting is wrong—only that its derivation is uncertain.

cells from a  $5 \times 5$  grid (18 pentominoes, 7 tetrominoes, 2 trominoes, 1 domino, 1 monomino; see Figure 41 in the Appendix).<sup>18</sup> Furthermore, any combination<sup>19</sup> of those shapes could be presented in the environment’s left and right panels (see Figure 31), rather than being limited to reflected or non-reflected versions of the *same* shape. The rationale for this change was to assess whether the software agent (described in the next section) could distinguish between shapes that highly resemble each other, but are, in fact, different shapes. For example, polyominoes 2, 3, 9, and 24 (see Figure 41 in the Appendix) are very similar: they only differ in their number of components, not in their overall shape or the relationships between their parts.

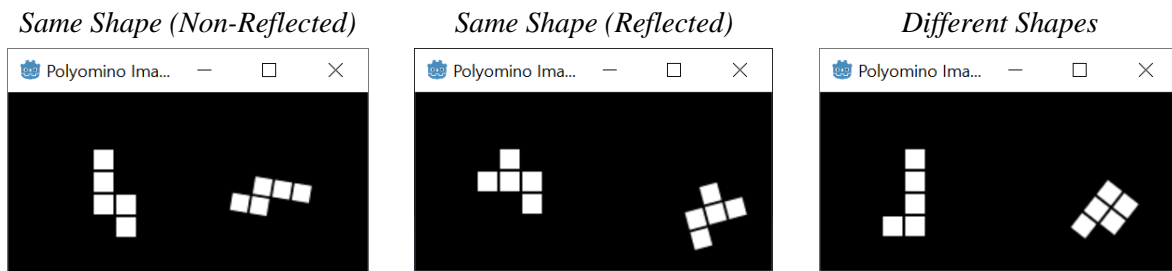


Figure 31. Categories of sensory stimuli pairs. The *same shape* may be presented in both panels (rotations, vertical and horizontal translations, and scaling allowed). The *same shape and its reflection* may be presented (rotations, vertical and horizontal translations, and scaling allowed). *Different shapes* may be presented in each panel (rotations, vertical and horizontal translations, scaling, and reflections, allowed).

In addition to increasing the number of supported shapes (from 12 to 29), new shape transformations—vertical and horizontal translations, and zooming in and out—were added to further perturb the shapes in the right panel from their reference (upright) shapes in the left

<sup>18</sup> Technically, these counts refer to the number of “one-sided” polyominoes. If the mirror images of a polyomino cannot be rotated onto one another—so that they completely overlap—then they are considered distinct one-sided polyominoes.

<sup>19</sup> While the environment could present any pair of shapes, the environment’s pair selection was constrained such that half of those pairs would show the same (non-reflected) shape in both panels and the other half a different shape.

panel. In total, the environment supports the following shape perturbations: rotations<sup>20</sup> in five-degree increments, horizontal and vertical shifts<sup>21</sup> in four-pixel increments, and nine different scales. Therefore, a (conservative) lower bound on the number of possible environmental states is over 40 million, with the right panel displaying any of approximately 1.4 million possible pixel patterns.

### ***Agent***

A LIDA-based agent was designed and partially implemented for the environment presented in the previous section. A high-level overview of its modules and processes is depicted in Figure 32. Each is summarized in the sub-sections below.

An important aspect of this LIDA agent is that its capacity to perform mental imagery operations is developed through environmental interactions. The perceptual and procedural knowledge resulting from these interactions is used as the basis for internal re-enactments that allow the agent to mentally simulate the effects of its actions on the world. This simulation-based LIDA agent is also intended to demonstrate that an agent's perceptual system can be enhanced via mental imagery. In other words, mental simulation can be an “epistemic process” that supports the non-symbolic generation of new knowledge.

---

<sup>20</sup> Note that the number of distinct rotations that are possible for each shape differs due to object symmetries. For example, shape 13 has 72 distinct rotational orientations (see Figure 42 in the Appendix), while shape 1 only has 18.

<sup>21</sup> The number of possible vertical and horizontal translations supported for each shape is different depending on their size and composition—ranging from 3 to 28 shifts in each direction); however, the majority of shapes support *at least* 10 translations in each direction, with many supporting over 20.

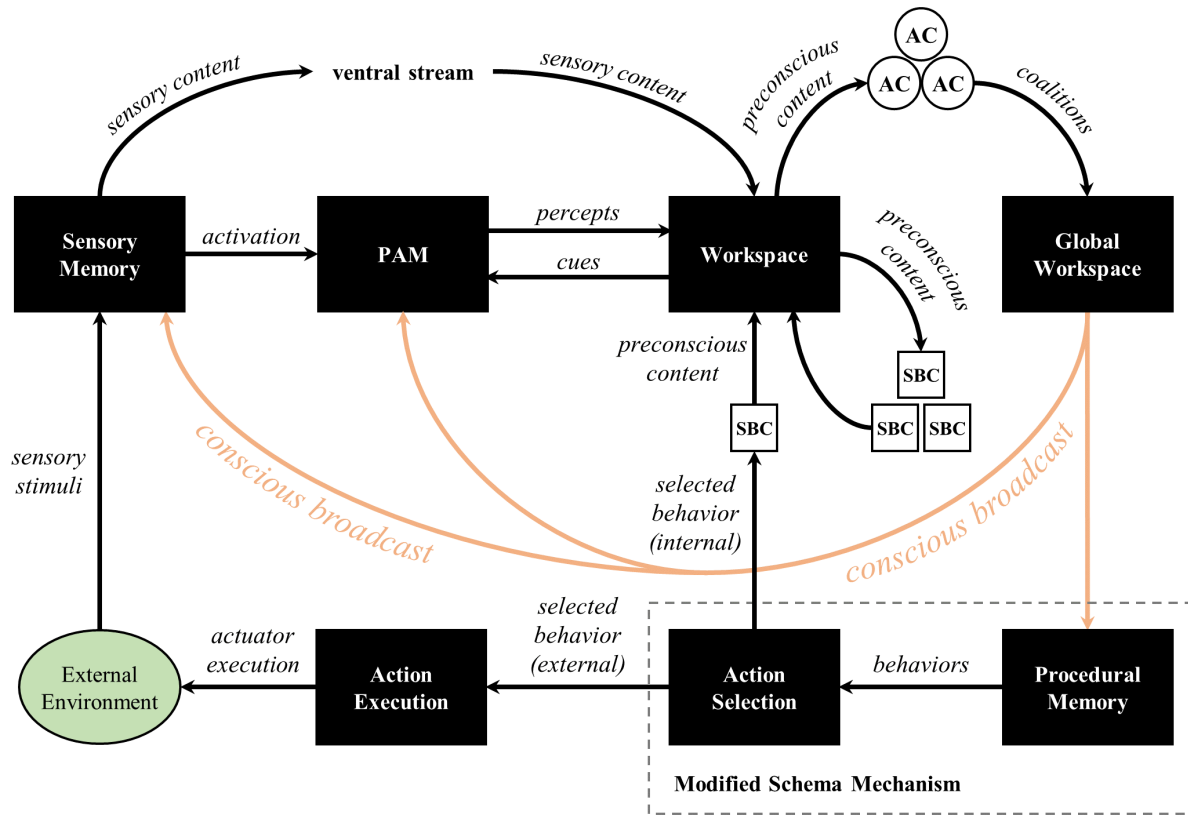


Figure 32. Cognitive cycle diagram for a LIDA-based agent. Both internally and externally executed behaviors are depicted. The agent's Procedural Memory and Action Selection modules were based on the schema-mechanism-based implementations described in Chapter 6.

To imitate a developmental period, the LIDA agent described in this section begins its life with an initial “play” phase, during which it is able to directly manipulate the shapes in the right panel. During this phase, the agent was capable of

- (1) rotating shapes (clockwise or counterclockwise),
- (2) translating shapes (vertically or horizontally),
- (3) moving shapes (closer or further away), and
- (4) advancing to the next shape.



Following the play phase, the agent lost the ability to move the pieces in the right panel. Its only available (external) actions were to decide—based on mental simulation and perceptual processes—whether the shapes presented in the environment’s left and right panels were the same or different.

**Sensory Memory.** Sensory Memory was implemented using a convolutional neural network (CNN; see Krizhevsky et al., 2012; LeCun & Bengio, 1995) based  $\beta$ -VAE (Higgins et al., 2017).<sup>22</sup> Recall (from Chapter 5) that variational autoencoders (VAEs; Kingma & Welling, 2013) are generative connectionist architectures composed of an encoder (or “recognition”) network and a decoder (or “generative”) network. The encoder network was used to implement Sensory Memory’s low-level visual feature detectors. The decoder network was used to support the generation of mental simulations (see Chapter 6, Simulator Structure Building Codelets).

On receiving visual sensory stimuli (i.e., two  $128 \times 128$  pixel, black-and-white images), those images were integrated into a simple visual sensory scene containing a pixel layer segregated into two panels. The VAE’s encoder<sup>23</sup> network was then used to *separately* generate sensory representations (i.e., modal probability distributions) corresponding to the content in each panel within the pixel layer. These sensory representations were then sent to the agent’s Current Situational Model (over LIDA’s “ventral stream”).

---

<sup>22</sup> The  $\beta$ -VAE’s overall architecture is summarized in Figure 38 of the Appendix, and its encoder and decoder network architectures are summarized in Figure 39 and Figure 40, respectively.

<sup>23</sup> The  $\beta$ -VAE’s encoder network (see Figure 39 in the Appendix) was composed of four alternating convolutional and max pooling layers.  $3 \times 3$  filters were used for all convolutional layers, with the number of filters per layer varying from 32 to 256.

**Perceptual Associative Memory.** Perceptual Associative Memory (PAM) was implemented as a content-addressable activation graph (see Chapter 5). This graph contained learned grounded concept representations for shapes in various orientations, as well as several built-in feature detectors. Built-in feature detectors<sup>24</sup> included “same,” “different,” “left,” and “right” amodal nodes, and a feeling node that quantified an agent’s “certainty” (positive valence sign). The “certainty” node was activated by a classifier structure building codelet (described later) based on content in the Current Situational Model.

Learned, elementary grounded concept representations (see Chapter 5) corresponding to encountered shapes were activated (i.e., received current activation) via their modal constituents (i.e., grounding modal nodes). Specifically, a modal node’s current activation is based on the scaled<sup>25</sup>, cosine similarity between its associated sensory representation and those for real or imagined sensory stimuli (in the Perceptual Scene). This current activation can then spread from modal nodes to their connected amodal nodes. If nodes receive sufficient activation, a reference to these node structures was added to the Current Situational Model (CSM)—i.e., they were instantiated into LIDA’s Workspace.

**Structure Building Codelets.** Several structure building codelets (SBCs) were implemented in support of perception and mental simulation. These included a sensory-integration SBC, a simulator SBC, a classifier SBC, and a timekeeper SBC.

---

<sup>24</sup> Ideally, the concepts of “same,” “different,” “left,” and “right” would be learned from experience. However, that level of conceptual generalization is well beyond the current implementation’s capabilities. As a result, structure building codelets were used to “cue” this content (i.e., activate it in a top-down fashion from the Current Situational Model).

<sup>25</sup> A sigmoidal activation function (see Figure 43 in the Appendix) was used to scale these cosine similarities to derive the modal nodes’ current activation values.

The sensory-integration SBC (cf. “multimodal-binding” SBC, Chapter 5) bound sensory representations—generated by Sensory Memory’s encoder network—to coordinating amodal nodes. The resulting node structures were then integrated into the agent’s Perceptual Scene—in association with either the left or right visual sensory scene elements. Conceptually, this can be viewed as linking amodal node structures to their corresponding locations in the agent’s visual field using “place nodes” (Madl et al., 2016).<sup>26</sup>

The simulator SBC for this agent was limited to fulfilling *action-based* mental simulations (i.e., spontaneous/involuntary mental simulations were not implemented). Whenever a behavior was selected for internal execution by Action Selection, its result was mentally simulated (via the  $\beta$ -VAEs generative network) in the Perceptual Scene’s right panel. In particular, the simulator SBC located the grounding sensory representation (modal probability distribution) for this node structure by iterating backward over (referential) activation links in a top-down fashion. Once located, the modal probability distribution was first passed through a *sampler*, which selected a *single* 16-dimensional latent vector from that probability distribution. This latent representation was then passed through the  $\beta$ -VAEs decoder network to generate a mental simulation of that scene element. (Figure 34 demonstrates the quality and variability of simulations for several shapes.) The simulator SBC then integrated this node structure and its associated mental simulation into the Perceptual Scene. This operation was performed by adding

---

<sup>26</sup> Even if bottom-up perception resulted in the instantiation of a PAM node (i.e., the sensory content in the visual sensory scene was “recognized”), these real and virtual sensory representations were still associated with a unique amodal node that characterized that sensory experience. This supported the agent’s ability to differentiate between multiple, identical objects in the same sensory scene (among other things). (This operation corresponds to Step (3) of the “multimodal perception” process described in Chapter 5.)

the node structure to the Perceptual Scene's node layer in association with the right panel, and its mental simulation was added to the visual sensory scene's right panel.

The classifier SBC monitored the node layer of the agent's Perceptual Scene. If the sensory content in the left and right panels were associated with the same shape (based on a comparison of their associated sensory content), then the classifier SBC created a "same" node in the Current Situational Model, which was then linked to the elementary grounded concept representations for those shapes in the Perceptual Scene. The "certainty" feeling node was also activated (given current activation), based on the scaled, cosine similarities between those shape's sensory representations.<sup>27</sup>

Finally, a timekeeper SBC was used in support of deliberation. In particular, each time the agent executed an action to advance to the next set of shapes in the environment, it entered into a deliberative mode of action selection. In particular, a "different shape" proposal was immediately generated on advancing the environment to the next shape, and a deliberation *timer* was started by a timekeeper SBC. If the timer expired before a "same" node was generated by the classifier SBC—which is functioning as an objector from James's ideomotor theory (see Franklin et al., 2016, sec. on Volitional decision making)—then the agent decided that the shapes were different. Otherwise, the "same" node might be consciously broadcast and used to instantiate an appropriate overt behavior for indicating that the shapes are the same.

**Attention Codelets.** Two attention codelets were implemented. The first was a general-purpose, activation-based attention codelet (referred to as a "default attention codelet" by Franklin et al.,

---

<sup>27</sup> Ideally, this operation would have been based on orienting behaviors and image introspection.

2016, p. 118). This attention codelet sought content in the Current Situational Model with the highest total activations. The second was a reconstruction-error-based attention codelet, that sought content associated with mental simulations that had high reconstruction errors.

**Procedural Memory and Action Selection.** Procedural Memory and Action Selection were implemented based on an enhanced, LIDA-compatible version of Drescher’s schema mechanism (see Chapter 6). Procedural Memory was initialized with built-in “bare” schemes (see Chapter 6) for each of the agent’s primitive actions: clockwise and counterclockwise rotation, zooming in and out, vertical and horizontal translations, indicating “same” shape, and indicating “different shape). From these, all other schemes were learned. Action Selection’s internal vs. external behavior determination logic was simplified to a simple toggle: rotations, zooming, and translations were externally executed in “play mode” and internally executed otherwise. Same/different shape actions were always externally executed.

### ***Results***

As I previously stated, this design was only partially implemented. Sensory Memory and Perceptual Associated Memory were largely implemented, along with a classifier structure building codelet and a simulator structure building codelet; however, these were never fully integrated with the schema-mechanism-based Procedural Memory and Action Selection implementations developed in Chapter 6. Complicating factors included:

- (1) the lack of a generalization process combined with a missing comprehensive decay strategy made learning too prolific to be scalable,
- (2) an unclear strategy for integrating incoming sensory stimuli with imagined sensory stimuli,

(3) the limited scalability of Procedural Memory’s composite actions (i.e., behavior streams).

Each of these items could be addressed in the future, but they require a prohibitive amount of additional conceptual and computational work to be included in this manuscript.

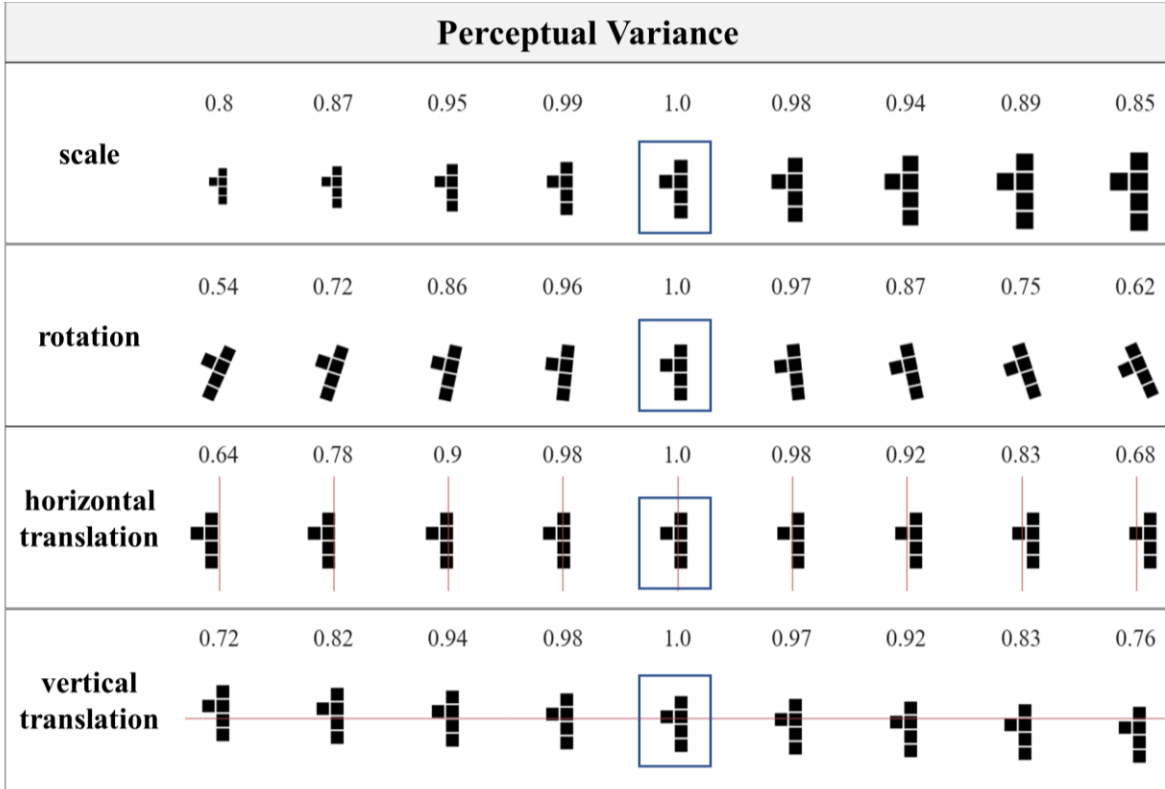


Figure 33. Perceptual variance. Polyominos surrounded by blue bounding boxes depict the reference image that other images are compared. The numbers above polyominos depict the cosine similarities of their corresponding sensory representations (i.e.,  $\beta$ -VAE generated modal probability distributions).

One interesting finding was that the convolutional neural network (CNN) based implementation of Sensory Memory is relatively sensitive to changes in the scale, rotation, and position of objects (see Figure 33). For example, a rotation difference of 10 degrees was sufficient to drop that shape’s node’s current activation below the Perceptual Associative Memory (PAM) instantiation threshold. That is, it would no longer be recognized as the same

shape. Consequently, mental simulations (specifically, action-based mental transformations) appear to be crucial for enabling this agent to perform its perceptual task.

In general, the  $\beta$ -VAE based simulator SBCs worked well for this environment. Mental simulations also exhibited the expected perceptibility (i.e., ability to be recognized by PAM) and variability (see Figure 34). The activation of PAM using sensory representations also generally worked well (see Figure 44 in the Appendix). Furthermore, modal nodes (see Chapter 5) for shapes that resembled one another received some degree of joint activation when those shapes occurred in the Perceptual Scene—for example, notice the joint activations for shapes 3, 9, and 24 in Figure 44 in the Appendix.<sup>28</sup> This suggests that more sophisticated recognition tasks may require a part-by-part inspection of similar objects to determine their identities.














































Simulation Quality and Variability								
								
								
								
								
								
<i>originals</i>	<i>simulations</i>							

Figure 34. Simulation quality and variability. Mental simulations were generated for learned sensory representations corresponding to the “original” images on the left.

<sup>28</sup> Figure 41 of the Appendix contains all polyomino shapes and their identifying numbers, for reference.

As a final note, the  $\beta$ -VAE's latent space appears to have learned some degree of “disentanglement” (see Chapter 5), though it can only be used in a very limited way to extrapolate beyond seen shapes. Figure 45 shows the simulations corresponding to sensory representations that lie on the same hyperplane as two reference shapes. Stepping between those shapes on the hyperplane (i.e., interpolation) worked fairly well—i.e., the expected perturbations (zooming, rotation, and translation) were apparent. However, the integrity of the shapes was eventually compromised when moving further away from those points on a hyperplane (i.e., extrapolation).

### **Related Work**

There have been several attempts at implementing mental imagery within cognitive architectures. I review two of these below. Shanahan's architecture was chosen because it is likely the only other Global Workspace Theory (GWT) based cognitive architecture that implements mental simulations. Soar was chosen because it provides a nice counterpoint to the implementation developed here.

#### ***Shanahan's Architecture***

Shanahan (2006) proposed a *brain-inspired* cognitive architecture that implements internal simulations, analogical representations (via topographically organized maps of neurons), and portions of Global Workspace Theory (GWT; Baars, 1988). Shanahan's architecture combines a low-level, *reactive*, behavioral system with a higher-level, *predictive* system that “simulates” action outcomes. These simulations can elicit affective responses (i.e., “emotions”) and guide action selection.



Shanahan's action selection is based on the "salience" of its actions, where possible actions are recommended by its low-level, reactive system. If one of the recommended actions is sufficiently salient, then it will be directly executed by its low-level system; otherwise, the higher-level predictive system will intercept its execution simulate its outcome. If the predicted outcome produces a sufficiently high affective response (i.e., emotional response), the salience of its corresponding action may be sufficient for enough to be executed. If not, alternate actions will be simulated until one has sufficient salience for execution.

Shanahan's architecture has many similarities to the LIDA-based implementation proposed here. Some of these are listed below:

1. It is based on the Global Workspace Theory (GWT) of consciousness.
2. It features internal (mental) simulations.
3. It uses analogical (iconic) representations.
4. It implements affective appraisals that can be used to evaluate the desirability of simulated outcomes.
5. It has a low-level reactive system that is roughly analogous to LIDA's Sensory Motor System.
6. It has a high-level predictive system that resembles the iterative execution of LIDA's internal behaviors combined with perceptual cueing for affective appraisals of their expected results.

However, there are also many differences (see Table 3).

Table 3. Comparison between Shanahan’s cognitive architecture and LIDA.

<b>Shanahan’s Architecture</b>	<b>LIDA</b>
purely <i>connectionist</i> architecture, based on “weightless neural networks” (Aleksander et al., 2009)	hybrid, <i>neuro-symbolic</i> architecture
learning is <i>supervised</i> —e.g., based on a predefined training script (Shanahan, 2006, p. 445)	learning is <i>unsupervised</i>
utilizes analogical, non-symbolic representations	utilizes analogical, non-symbolic representations <i>and</i> non-analogical, symbolic representations
single sensory modality (i.e., vision)	multi-modal
potential actions are recommended by a reactive <i>online</i> system—based on current sensory inputs (i.e., “real” inputs)	potential actions (i.e., behaviors) are recommended by an <i>offline</i> cognitive system (via Procedural Memory)—based on an internal model of the current situation that contains both “real” and “virtual” sensory content
mental simulations are always directly related to current action possibilities (i.e., proximal intentions)	mental simulations can be arbitrarily detached (spatially and temporally) from an agent’s immediate concerns
does not distinguish between unconscious and conscious mental simulations	models preconscious, never conscious, and conscious mental simulations
learning is based on direct associations between neural areas, bypassing the global broadcast—i.e., learning is largely (exclusively?) unconscious	<i>all</i> learning is based on the global broadcast—i.e., learning requires consciousness

Shanahan’s architecture is elegant, but it is relatively limited in the cognitive phenomena it can model. Mental simulation can be seen as a minimal, predictive (offline) extension to a reactive (online) behavioral core, and this capability is only employed when the consequences of its current actions lack sufficient salience (i.e., desirability). In particular, Shanahan’s implementation of mental simulation is confined to the prediction of action consequences, and those predictions are largely tethered to an agent’s immediate environmental conditions (i.e., its

current place and time). As such, Shanahan's architecture has no ability to imagine counterfactuals or to anticipate distal events: mental simulations of distant places or points in time are impossible. Moreover, Shanahan does not attempt to utilize mental imagery for perceptual discrimination or creative tasks. Imagery is solely used to assess the salience (desirability) of action consequences.

Shanahan's architecture has a very basic internal environment with a single sensory modality, and it makes no attempt to integrate "real" sensory content with "virtual" sensory content. Its implementations of long-term memory (via a long-term visual buffer) and learning (based on a predefined script) are also very limited. While Shanahan's information processing features a global broadcast that was inspired by Baar's (1988) Global Workspace Theory, it does not appear to distinguish between unconscious, pre-conscious, and never conscious representations and mental simulations.

The LIDA-based implementation developed throughout this manuscript does not suffer from these limitations. Therefore, it can be used to model a much larger cross-section of cognitive phenomena than Shanahan's architecture. That said, LIDA is also exceedingly complicated, and its computational implementation is far from complete.

### *Soar*

The Soar cognitive architecture (Laird, 2012) has its roots in the early work of Allen Newell and Herbert Simon, particularly the Logic Theorist (Newell & Simon, 1956) and the General Problem Solver (Ernst & Newell, 1969). Subsequent work by John Laird and Paul Rosenbloom expanded on these methods to create a general-purpose cognitive architecture that is less problem specific than its predecessors.

For much of its history, Soar was a purely *symbolic* cognitive architecture. However, more recently, *non-symbolic* processes have been added (Lathrop & Laird, 2007; Wintermute, 2012). Laird describes Soar’s non-symbolic processes (e.g., mental imagery) as “co-symbolic” (Laird, 2012, p. 21); that is, they are processes that manipulate non-symbolic representations in service of Soar’s symbolic decision-making and reasoning algorithms (e.g., means-ends analysis, backward chaining, and operator subgoalting; see Laird, 2012).

Wintermute (2012) detailed Soar’s most recent implementation of mental imagery. Like Shanahan’s architecture, this work focused exclusively on the internal simulation of action consequences, which Wintermute referred to as “simulative imagery” (Wintermute, 2012, p. 3). Wintermute was also primarily concerned with imagery for spatial tasks, though the same basic design could be extended to other problem domains.

Soar supports several imagery operations that are functionally analogous to Kosslyn’s image generation and image transformation (described earlier in this chapter). Image generation is implemented using “memory retrieval” and “predicate projection” (see Wintermute, 2012, p. 12). Image transformation is implemented using special-purpose, continuous action controllers (e.g., navigation planners). These controllers often support both the execution and simulation of actions (see Wintermute, 2012, p. 12). (Soar does not support functional equivalents to Kosslyn’s image introspection or image maintenance.)

Once Soar’s imagery processes generate and manipulate their non-symbolic representations, *high-level perception* then maps them to symbolic representations (e.g., predicates). Prior to Wintermute’s (2012) implementation of mental imagery, there was no need to distinguish between “low-level” and “high-level” perception in Soar. Environmental stimuli

were transduced directly into symbolic representations, without intervening non-symbolic representations. After introducing mental imagery, there was a need to specify both low-level and high-level perceptual operations in Soar. Low-level perception transforms raw environmental states (sensory stimuli) into non-symbolic representations. High-level perception then transforms non-symbolic representations into symbolic representations (see Figure 35, Left Panel). Soar's actions were similarly split into low-level and high-level actions.

Figure 35 contains a side-by-side comparison of Soar (left panel) and LIDA (right panel) using the functional and representational terminology established by Wintermute (2012). As this diagram demonstrates, both architectures support roughly comparable functional components; however, the interactions between these components are often quite different. Soar and LIDA also use different representational formats. Where LIDA's representations are generally hybrid (symbolic/non-symbolic) and fully integrated (depicted as  $R_a/R_c$ ), Soar's representations are segregated into symbolic ( $R_a$ ) or non-symbolic ( $R_c$ ) representations.

To delve deeper into these architectural differences, Wintermute's functional designations were mapped to their closest corresponding LIDA modules and processes. These are depicted in Figure 36. *Low-level perception* corresponds to the activation of low-level feature detectors in LIDA's Sensory Memory and the subsequent generation of sensory representations. *High-level perception* corresponds to the bottom-up activation and cueing of Perceptual Associative Memory (PAM), and the instantiation of any resulting percepts. *Low-level action* corresponds to the execution of motor plans by Action Execution—i.e., sending motor commands to an agent's actuators. And *high-level action* corresponds to the instantiation and

selection of behaviors (instantiated schemes) by LIDA's Procedural Memory and Action Selection modules, respectively.

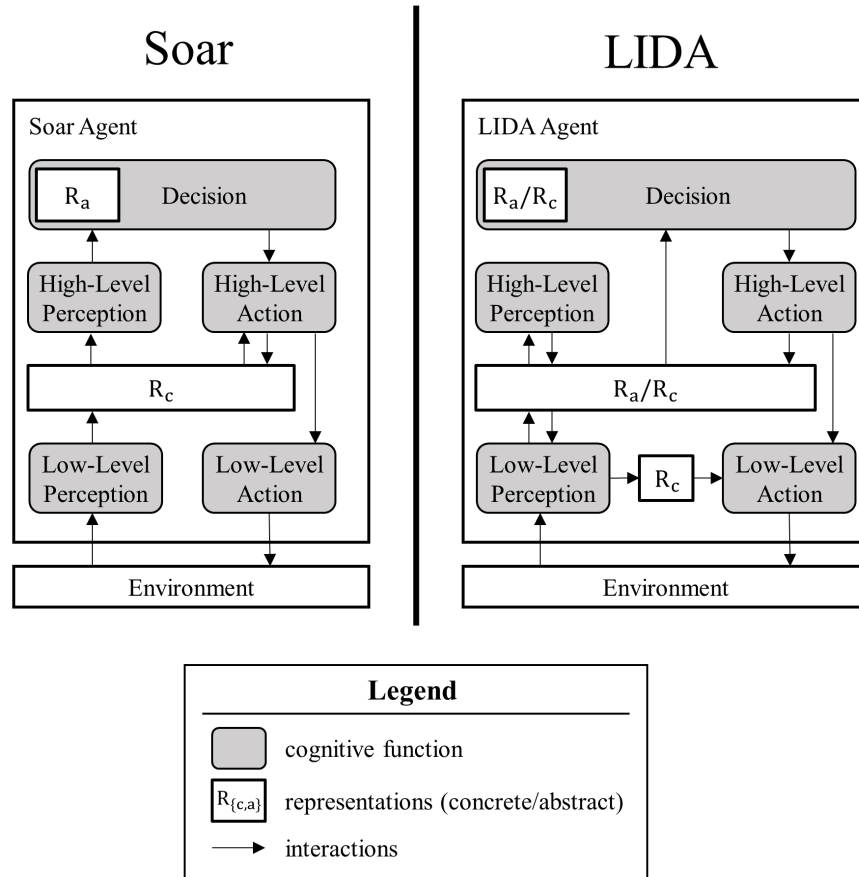


Figure 35. High-level comparison of Soar's and LIDA's mental imagery implementations. The left panel shows Soar's implementation based on Fig. 7(a) from Wintermute (2012). The right panel shows LIDA's implementation in a similar style.  $R_a$  and  $R_c$  denote "abstract" and "concrete" representations respectively. These terms are used by Wintermute (2012) instead of symbolic and non-symbolic representations, though their usages appear to be functionally equivalent.

*Decision* is the most challenging functional component to map, as LIDA's decision-making processes are more diffuse and subtle than Soar's. To a first approximation, Soar's decision processes likely correspond to the following components in LIDA: (1) preconscious activity in LIDA's Current Situational Model (e.g., the cueing of long-term memory modules and the creation of preconscious content by structure building codelets), (2) the attentional processes

that operate on those preconscious representations (i.e., attention codelets, the competition in the Global Workspace, and “orienting” behaviors), (3) Action Selection’s choice of a behavior for internal or external execution, and (4) the “epistemic” behaviors<sup>29</sup> that support perception and simulation-based offline cognition.

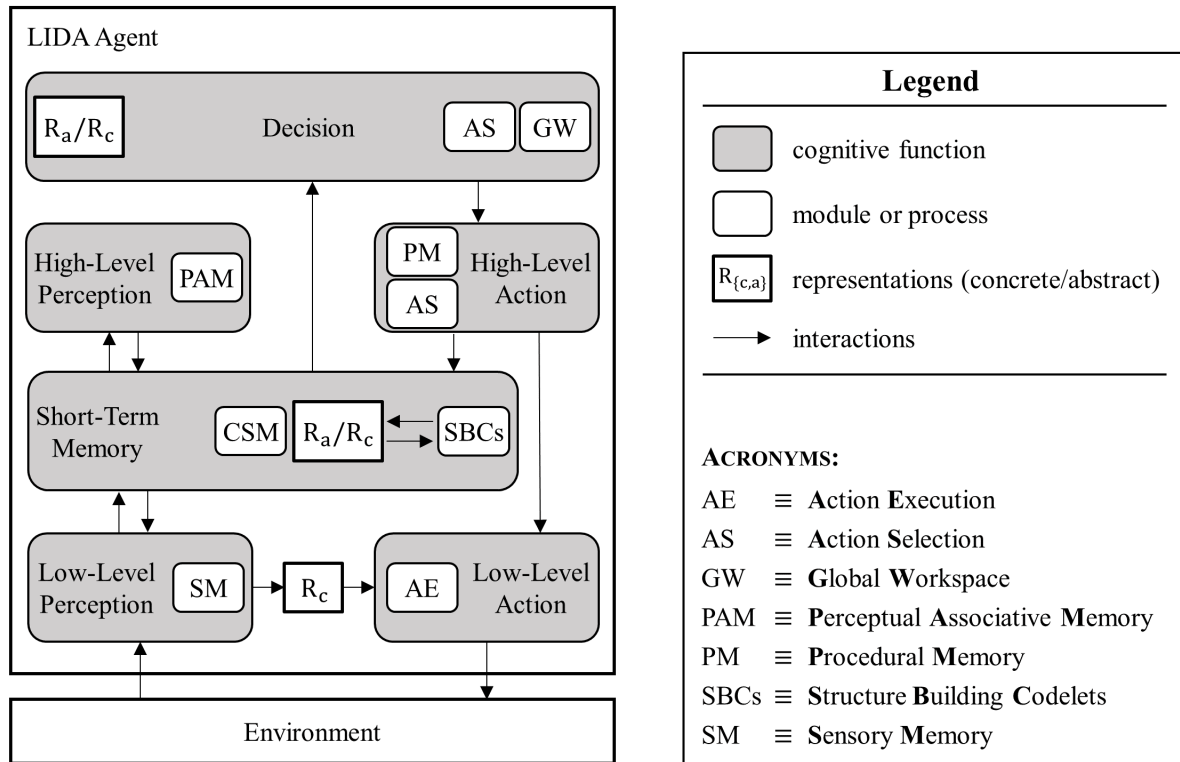


Figure 36. Functional mapping of LIDA's module and processes for comparison with Soar. Recall that  $R_a$  and  $R_c$  stand for “abstract” (symbolic) and “concrete” (non-symbolic) representations, respectively.  $R_a/R_c$  stands for a hybrid representation that integrates both.

While some of the differences between LIDA and Soar are quite subtle, there are some fundamental differences between these architectures that manifest in their implementations of mental imagery. Table 4 summarizes many of these differences.

<sup>29</sup> Epistemic behaviors (cf. “epistemic actions”; see Kirsh & Maglio, 1994) might include internally executed epistemic behaviors (e.g., image transformations) or externally executed epistemic behaviors (e.g., moving a chess piece on the board to visualize that new position).

Table 4. Comparison between Soar and LIDA.

<b>Soar</b>	<b>LIDA</b>
Decision processes are based on rule-based symbolic manipulation.	Decision processes are based on imagistic and perceptual processes that operate on hybrid (symbolic/non-symbolic) representations.
Imagery is a <i>built-in</i> capability based on the execution of special-purpose continuous action controllers in “imagery mode.”	Imagery is a <i>learned</i> capability based on the internal (covert) execution of behaviors that were acquired through the internalization of external (overt) action consequences.
Imagery is <i>initiated</i> by the execution of high-level (abstract) actions, but subsequent imagistic operations (e.g., transformations) are fulfilled by built-in continuous action controllers.	Volitional mental imagery is based on the <i>iterative</i> , often multi-cyclic, execution of internal behaviors.
Supports a single sensory modality (i.e., spatial)	Supports multi-modal mental simulations
Imagery is based on purely non-symbolic representations.	Imagery is based on hybrid (symbolic/non-symbolic) representations.
Supports the simulation of action consequences (i.e., “simulative imagery”)	Supports the simulation of action consequences <i>and</i> non-action-based mental simulations (e.g., the mental simulation of objects, events, situations, etc.)
Imagery is optional. It is only needed when agents are faced with tasks that require highly detailed representations.	Imagery is pervasive. Perception is continually supported by mental simulations. Conscious thought is largely imagistic.
Imagery’s primary function is to augment abstract (symbolic) representations with additional details.	Imagery is a core function of simulation-based cognition. It supports offline cognition via experience-based predictions, image transformations, and introspection (among other things).

As with Shanahan’s architecture, Soar is more focused on solving fundamental engineering problems, whereas LIDA is more focused on cognitive modeling. As a result, LIDA can conceptually model more mental-imagery-related cognitive phenomena than Soar—e.g., spontaneous and intentional mental imagery, image inspection and image maintenance, conscious vs. unconscious mental simulations, multi-modal perception and simulation, etc. LIDA



also models and implements how procedural and perceptual learning support the capacity to mentally simulate one's action and their consequences. By comparison, Soar largely delegates its imagistic operations to built-in continuous-action controllers.

Perhaps the largest difference between these architectures relates to their views on the role and importance of mental imagery in supporting offline cognition. Wintermute (2012) contended that imagery is primarily needed to compensate for the informational differences between non-symbolic and symbolic representations. According to Wintermute, environments must be internally represented at variable levels of detail—balancing precision vs. abstraction—and he viewed imagery as the mechanism by which the more detailed representations are manipulated. In contrast, the implementation of LIDA proposed here uses mental imagery as the primary mechanism for generated new knowledge, rather than rule-based symbolic manipulations. Thought and decision-making in the LIDA implementation developed here is inherently imagistic and perceptual, not symbolic. And, in this view, mental simulation is foundational to offline cognition.

As a final note, supporting “simulative” mental imagery required considerable architectural changes in Soar. Perception and action had to be split into high-level and low-level components, and a completely new class of representations (i.e., non-symbolic representations) needed to be supported. By contrast, LIDA's more full-featured implementation of mental imagery was largely enabled by expanding on existing architectural components.

## **Discussion**

Embodied, simulation-based cognition depends on the coordinated application of imagistic and perceptual processes. These “epistemic devices” (Fisher, 2006) generate, and make available,

*new knowledge* by allowing individuals to “simulate reality at will” (Moulton & Kosslyn, 2009). Past events can be recreated, future events anticipated, and current events augmented (e.g., based on their most probable causes; Albright, 2012). These epistemic processes operate both consciously (mental imagery) and unconsciously (mental simulation) in support of offline and online cognition.

Mental imagery is an exceedingly complex and multi-faceted cognitive ability, and many of its aspects remain poorly understood. However, a core theme has emerged throughout this thesis is that cognitive is not only in service of action (Franklin, 1995; M. Wilson, 2002), it is substantially *action-based*; that is, cognition *is* action. Simulation-based reasoning is, by and large, an *active* process, based on the internal (covert) execution of behaviors. This internal execution of behaviors controls multi-part image generation, image transformation, image inspection, and modulates the attentional components of image maintenance.

The LIDA implementation presented here predicts that mental imagery, mental simulation, and active perception crucially depends on consciously mediated, volitional, and perhaps even automatized forms of action selection. Furthermore, mental imagery develops as a skill, based on the internalization of environmental interactions that manifest through the acquisition of procedural and perceptual knowledge.

## Chapter 8

### Closing Remarks

Theories of embodied, simulation-based cognition solve what Pezzulo and Castelfranchi (2007) called the *symbol detachment problem*. These theories explain

- (1) how representations and processes can be decoupled (temporally, spatially, or otherwise) from a cognitive system's immediate inputs (sensory stimuli) and outputs (motor commands), and
- (2) how these “detached” representations and processes can retain their grounding, intentionality, and embodiment.

ES-Hybrid—the cognitive theory developed throughout this manuscript—falls within this tradition.

### Contributions

Like most cognitive theories, ES-Hybrid developed in response to specific shortcomings in its predecessors (e.g., perceptual symbol systems; Barsalou, 1999). In particular, ES-Hybrid attempts to provide a unified account of concrete and abstract concept representation, and to be more amenable to computational implementations and lower-level conceptual models (such as those that were developed within the LIDA cognitive architecture).

While ES-Hybrid models both *grounded* and *ungrounded* cognition, it is, by and large, a theory of *grounded* cognition. However, rather than assuming that grounding is a compulsory property of mental representations (cf. perceptual symbol systems), it assumes that grounding is a desirable property that needs to be established through experience and speculation. In this

conception, grounding is not a singular event but, rather, the gradual acquisition of information-bearing associations with one's (internal or external) environment. Some representations—in particular, those created by predictive processes to represent hypothetical, unknown referents—are *initially* ungrounded. And yet, those representations are almost always *eventually* grounded.<sup>30</sup>

A fundamental assumption of ES-Hybrid is that cognitive systems make extensive use of amodal representations (i.e., symbolic representations). These amodal symbols have their closest biological analogs in Damasio's (1989) convergence zones and the trans-modal "hubs" that are used in hub-and-spoke models. That is, they are *coordinating* representations that "point to" where information is located, without directly encoding that content.<sup>31</sup> Crucially, the use of amodal representations does not necessitate the use of explicit, rule-based, symbolic manipulations.

According to ES-Hybrid, *all* mental representations are best viewed as composite modal/amodal structures. In this view, I am in agreement with Michel (2021) who argued that the modal/amodal dichotomy is best viewed as a graded property—i.e., representational structures can be characterized as points on a spectrum between purely modal and purely amodal.

---

<sup>30</sup> While I did not discuss this possibility earlier in the manuscript, it is likely that representations could lose their grounding—for example, as a result of interference, decay, or damage to the nervous system. Therefore, grounding may be an impermanent property of representations that must be actively maintained.

<sup>31</sup> Compare this with "cross-modal conjunctive representations" (CCRs; Binder, 2016) and "compressed multimodal representations" (Barsalou, 2016a), which do encode modal information.

In addition to developing a new hybrid theory of embodied, simulation-based cognition, this work made numerous contributions to the LIDA cognitive architecture. Many of these are listed in Table 5.

Table 5. Contributions to the LIDA cognitive architecture.

	Contribution	Chapter
1	multimodal perception	5
2	multimodal perceptual learning	5
3	sensory representations	5
4	grounded concept representations	5
5	multimodal mental simulation	5
6	simulation-based attention	5
7	cognitive “object” maps	5
8	“instructionist” procedural learning	6
9	motor cognition	6
10	behavior streams	6
11	action selection → internal vs. external	6
12	mental imagery → image generation	7
13	mental imagery → image transformation	7
14	mental imagery → image inspection	7
15	mental imagery → image maintenance	7

One strength of my implementation is that it did not require any major modifications to LIDA’s conceptual model. I added no new modules or submodules, and I only required a few modifications/expansions to LIDA’s existing modules, processes, and representational system. The most significant updates included

- (1) subtyping nodes into *modal* and *amodal* nodes (see Chapter 5),
- (2) subtyping activation links into *referential* and *non-referential* links (see Chapter 5),
- (3) adding new structure building codelets (e.g., simulator and multimodal-binding SBCs; see Chapter 5),

- (4) adding new attention codelets (e.g., reconstruction-loss attention codelets; see Chapter 5), and
- (5) updating Action Selection to support the general, internal (covert) execution of behaviors (see Chapter 6).

These minor alterations are especially striking when compared with Soar’s implementation of “simulative” mental imagery (Wintermute, 2012), which resulted in a major overhaul of its perception and action modules (into new high-level and low-level modules), new representational formats (i.e., “concrete” non-symbolic representations), new processes to manipulate them (i.e., continuous-action controllers), and new modes of action execution (i.e., “execution mode” vs. “imagery mode”).

Another strength of my implementation is that it is largely based on well-developed and well-analyzed “off-the-shelf” techniques in machine learning (e.g.,  $\beta$ -VAEs). Consequently, there is a large community of researchers, developers, and special-purpose tools that facilitate their use. In general, the computational techniques used here were chosen because they were relatively well-understood and met the minimum requirements demanded by ES-Hybrid and LIDA. As more efficient or appropriate computational implementations are identified, they can and should replace the implementations used here.

## **Related Work**

Embodied, simulation-based cognition (ES) has rarely been attempted in a cognitive architecture. Shanahan’s (2006) architecture is arguably the closest to an ES cognitive architecture (see Chapter 7). Other cognitive architecture include aspects of mental simulation or mental imagery

(e.g., see Chella et al., 1997; Rosenbloom et al., 2016; Wintermute, 2012), but they also depend on symbolic offline reasoning systems; therefore, they are not ES-based cognitive systems.

There have been numerous attempts at implementing *portions* of a perceptual symbol system (PSS; Barsalou, 1999) within software; however, few have tried to systematically build a PSS from the ground up, based on first principles (e.g., perceptual symbols and simulators). Many implementations attempt to address topics of high theoretical interest, such as abstract concepts and language, without a firm implementation of PSS's basic components.

Arguably the best example was Joyce et al.'s (2003) attempt to create a PSS starting from basic foundational components (e.g., perceptual symbols). They implemented a connectionist, computational model based on a recurrent neural network architecture (see Chapter 2) that they call the Connectionist Perceptual Symbol System Network (CPSSN), and they applied it to labelled video sequences. The authors claimed that CPSSN is a mechanism for implementing perceptual symbols, and that it contains “categorical information summarising the event/episode.” The network also has some rudimentary simulation like capabilities. Unfortunately, their analysis makes no mention of concept representations, simulators, frames, or simulation-based cognitive processes, so it is difficult to assess the future prospects of this approach. Moreover, the  $\beta$ -VAE based implementation of sensory representations and elementary grounded concepts presented in Chapter 5 is preferable in almost every way for implemented generative, modal representations. The primary advantage the CPSSN has over the  $\beta$ -VAE used here is that it captures a temporal dimension; however, that could easily be addressed by using a different  $\beta$ -VAE architecture (e.g., a recurrent  $\beta$ -VAE architecture).

As another example, Perlovsky and Illin (2012) argued that computational accounts of PSS require new mathematical frameworks that are “different from traditional artificial intelligence, pattern recognition, or connectionist methods,” and they propose the use of Dynamic Logic (DL) for that purpose. They experimentally show that DL can implement object/situation representations and recognition and may be capable of supporting multiple modalities; however, the connections between DL’s operations and PSSs are highly speculative.

Finally, Stramandinoli et al. (2011) attempted to develop a mechanism for learning abstract concepts through sensorimotor experiences using a humanoid (iCub) robot. This work built on prior work by Cangelosi and Riga (2006) by extending its “higher order grounding phases” to include more abstract concepts. In the basic grounding phase, the robot learned a set of action primitives (e.g., PUSH, PULL, GRASP, STOP, and RELEASE) by observing and imitating a teacher along with their corresponding linguistic labels. In the first higher-order grounding phase, linguistic descriptions that contained multiple action primitives were presented to the robot, such as “KEEP [is] GRASP [and] STOP” (Stramandinoli et al., 2011, p. 470). And in the last higher-order grounding phase, meaning was transferred to abstract concepts, for example, “ACCEPT [is] KEEP [and] SMILE [and] STOP” (Stramandinoli et al., 2011, p. 471). Unfortunately, this work suffers from several issues. First, what is learned seems to be a form of analogical reasoning rather than the generalization of sensorimotor experiences into abstract concepts. Second, the control structures implemented for the iCub do not seem to be based on any of the fundamental components of a perceptual symbol system (e.g., perceptual symbols and simulators). Finally, I am skeptical that these combinations of action primitives (e.g., keep, smile, and stop) adequately capture the meanings of more abstract words (e.g., accept).



Leaving implementations of perceptual symbol systems behind and looking to the broader embodied, simulation-based cognition (ES) literature, Barsalou et al. (2008) proposed the Language and Situated Simulation (LASS) cognitive theory. This theory suggests that multiple systems support conceptual representation and processing. They focus on linguistic and simulation-based systems, though they allow for other systems, such as those based on distributional semantics (see Chapter 2). They argued that “deep” conceptual processing requires simulation, while the processing of linguistic systems tends to only support superficial conceptualization. However, depending on the type of task an individual is performing, one system may dominate the other in conceptual processing.

LASS is very similar to Paivio’s (1986/1990) dual coding theory (DCT), which proposes that cognitive systems are composed of two specialized subsystems for handling verbal and nonverbal information. Each subsystem is structurally and functionally distinct, allowing it to operate independently; however, they are also functionally interconnected, allowing activity in one system to initiate activity in the other.

The verbal subsystem is responsible for the perceptual and motor activities associated with verbal and written language. The nonverbal subsystem is responsible for handling the “sensory properties of things, relations among them, and their behavioral ‘affordances’” (Paivio, 1990), and is inherently “imagistic” (that is, its operations are based on mental imagery). These subsystems are further divided into modality-specific components, which can process information separately (unimodal representations), or integrated together (multimodal representations), depending on the needs of a task. There are no mediating (system-agnostic) representations that coordinate communication between the verbal and nonverbal systems.

The mental representations within each subsystem are referred to as *logogens* (verbal) and *imagens* (nonverbal). Logogens are “sequential structures of increasing length, from phonemes (or letters) to syllables, conventional words, fixed phrases, idioms, sentences, and longer discourse units—anything learned and remembered as an integrated language sequence.” Imagens are representations of “natural objects, holistic parts of objects, and natural groupings of objects” from which mental images are generated. Both logogens and imagens can be multimodal and hierarchical. Paivio notes that “logogens have no meaning in the semantic sense. They are directly ‘meaningful’ only in that they can be activated by stimuli similar to those involved in the original formation of the corresponding logogens” (Paivio, 2014, pp. 146–147). In contrast, imagens are intrinsically meaningful, and when they are activated their “imaginal memory traces resemble the perceived objects and scenes they represent” (Paivio, 2014, p. 147).

*Referential connections* (cf. ES-Hybrid’s referential connections) between systems allow cross-system activation and provide a mechanism for “objects to be named and names to activate images that represent world knowledge” (Paivio, 2014). Inter-logogen and inter-imagen *associative connections* (cf. ES-Hybrid’s non-referential connections) within each subsystem provide additional relational/contextual information about representational units.

DCT assumes that all mental representations are learned from “perceptual, motor, and affective” experience and that those representations “retain those experientially derived characteristics so that representational structures and processes are modality specific rather than amodal” (Paivio, 1986/1990, p. 55). Furthermore, DCT’s mental representations are never completely abstract: “the modality-specificity of logogens and images excludes abstract mental representations such as propositions. Thus the functional domains associated with stimulus

meaning and cognitive abilities are conceptualized entirely in terms of modality specific logogens and imagens” (Paivio, 2014, p. 146).

Barsalou et al. (2008) contend that while LASS is similar to DCT, it places less emphasis on its linguistic system and more emphasis on its simulation system (Barsalou et al., 2008, p. 253). ES-Hybrid diverges even further from DCT since ES-Hybrid uses coordinating amodal representations, and it allows for completely “abstract” (i.e., symbolic) representations. Another difference between ES-Hybrid and DCT (and LASS) is that its hybrid (modal/amodal) representations are processed by a single system rather than multiple, complementary, special-purpose systems. Nevertheless, DCT’s referential and associative connections are very similar to ES-Hybrid’s referential and non-referential associations.

Finally, Louwerse (2018) detailed a “unifying account” of symbolic and embodied cognition that combines the meaning derived from perceptual experiences with the meaning derived from linguistic contexts (e.g., distributional semantics; see Chapter 2). His theory is based, in part, on ideas from Deacon’s (1997) “hierarchy of signs”—which is based, in turn, on Peirce’s semiotics (see Chapter 2).

Deacon proposed that language processing could be explained using a hierarchy of iconic, indexical, and symbolic processes, where indexical associations can be used to give symbolic relationships meaning. Words derive their meaning through indexical relationships that are grounded in iconic relationships. Furthermore, Deacon argued that not all symbols need to be grounded in perceptual experience, but can, instead, get their meanings exclusively through indexical relationships with other symbols (cf. distributional semantics; Chapter 2).

A critical aspect of Louwerse's theory is that words for concrete and abstract concepts both rely on the same underlying mechanisms. Specifically, he stated that both rely on indexical relations, "but the extent to which abstract and concrete concept words rely on language statistics or perceptual simulations differs" (Louwerse, 2018, p. 584).

While ES-Hybrid does not make any explicit claims about language processing or the use of distributional semantics, it is not in opposition to Louwerse's (2018) ideas. ES-Hybrid also suggests that indexical associations are important for the derivation of meaning, but rather than using them as the basis for contextual statistics, it suggests that indices are used to drive grounding through active exploration and speculation. That is, indices provide clues to where to search for grounded meaning. Louwerse also acknowledges that symbols can be ungrounded, and that abstract concepts are ungroundable (Louwerse, 2018, p. 584), which is consistent with ES-Hybrid's account (see Chapter 3).

### **Directions for Future Work**

A conscious decision was made at the onset of this endeavor to favor breadth over depth. My objective was not only to add support for mental imagery and mental simulation in LIDA, but to show how embodied, simulation-based cognition could be used as a fundamental organizing principle within LIDA. As such, a coherent and relatively comprehensive "mid-level" theoretical picture seemed preferable to a detailed treatment of some specific cognitive process.

Consequently, there are more open issues, and opportunities for future work, than I could hope to enumerate here. Nevertheless, I will highlight a few immediate directions for further research.

### ***Referential and Non-Referential Associations***

The theory of grounding presented here depends on the establishment of “referential associations” (see Chapter 3). Referential associations establish correspondence-based relationships between representations and their referents, whereas non-referential associations broadly encompass all other types of associations (e.g., causal, indexical, analogical, and relational). Whether natural minds *learn* how to represent these associations through experience, or whether they are *innate* representational capabilities is a topic that needs to be explored.

As a computational simplification, one could assume that innate, cognitive processes exist for implementing these associative relationships, and this has been a tacit simplification that I have employed throughout this manuscript (with LIDA’s structure building codelets serving this function). However, this is clearly an oversimplification. Many non-referential associations appear to be grounded concepts in their own right (e.g., spatial and temporal relationships). Therefore, they should have their own grounding sensorimotor representations, and they should be simulatable. Referential associations, on the other hand, may be more fundamental, and perhaps different-in-kind, from non-referential associations. Consequently, it is possible that the cognitive processes that support the creation of referential associations are necessarily innate.

### ***Abstract Thought***

A basic assumption of the theory developed here is that offline cognition is based on imagistic and perceptual processes—not explicit, rule-based, symbolic manipulations. This is not a denial of the existence of abstract thought, but rather a belief that the same cognitive processes involved in simulation-based offline cognition can be used to realize more abstract, and algorithmic,

modes of reasoning. Substantiating this belief computationally should be a priority of future research.

Analogical/metaphorical thought (Lakoff & Johnson, 1980/2008) is clearly part of the solution. It is undoubtedly true that individuals often think about abstract concepts using deep or shallow metaphors with concrete concepts (“time flies like an arrow”). However, this is not the whole picture. Purely abstract thought (e.g., pure mathematics and formal symbolic logic) exists. What form does simulation-based thought take when modal simulations are impossible?

One intriguing possibility is that abstract thought may rely, in part, on “degenerate” modal simulation, where the node structures for ungrounded representations can be deployed in service of offline cognition. The mechanisms for these degenerate simulations may be largely identical to those used for full-blown modal simulations, but the top-down activation of concept representations during simulation necessarily terminates before reaching modal content. Such *amodal mental simulations* are not only possible, but the LIDA-based implementation developed here essentially predicts their existence. For example, many of the imagistic (epistemic) operations described in Chapter 7 will continue to function in the absence of full-blown simulations. In particular, internal (covert) actions could be used to volitionally control the introduction of amodal representations into LIDA’s Current Situational Model. Once introduced, non-referential associations between that amodal representation and other representations in long-term memory could be used to cue long-term memory and activate other situationally relevant representational structures.

## *Context and Contextual Semantics*

Mental simulations are context-sensitive (Barsalou, 2003, 2016b) and context can be instrumental in determining the meaning of things (e.g., distributional semantics; Baroni & Lenci, 2010; Erk, 2012; Landauer & Dumais, 1997; Lenci, 2008; Lund & Burgess, 1996). Moreover, symbolic and non-symbolic representations are often ambiguous in the absence of context. For example, the English word “chair” can refer to an object or a person; the word “gift” means something quite different in German than in English; and specific objects (e.g., to-go mugs at a coffee shop), events (e.g., commutes to work), and locations (e.g., parking spaces in a parking lot) are often too similar to designate (as a specific instance) without accompanying context (see Chapter 3, Designation and Disambiguation).

Of particular interest to the present work is determining how physical and situational contexts can be encoded using conceptual representations, and how this encoded context can guide and constrain active perception and conceptual learning. Recall that a central component of the theory presented here is that initially ungrounded representations are often generated by top-down, predictive processes. These ungrounded (amodal) representations are non-referentially associated with grounded (modal) representations that serve as contextual clues (or contextual cues; Chun & Jiang, 1998) to the identity of those unknown referents. These contextual clues can guide the selection of grounding epistemic actions. And, in general, I view contextual semantics as a means of facilitating grounding and extracting the most information from each grounded representation.

The mechanisms by which physical and situation context is represented, and its influence on various cognitive processes (e.g., perception, simulation, and action selection) need to be

explored. Furthermore, developing these theoretical elements is a prerequisite for modeling episodes and episodic memory systems, which are currently unspecified in ES-Hybrid and its LIDA implementations.

### ***Conceptual Generalization***

The current implementation lacks a mechanism for generalizing from elementary concept representations to more abstract representations (e.g., for objects and categories of objects). Elementary grounded concept representations function like Harnad's (1990) iconic representations, and an additional generalization process is needed to learn the equivalent of Harnad's (1990) categorical representations.

Recall from Chapter 2 that iconic representations are non-symbolic representations that encode the distinctive features of concept instances, while categorical representations are non-symbolic representations that capture their most important *invariant* features—i.e., the relative weighting of features based on their importance in determining category membership. For example, color may be irrelevant for determining that an object is a chair but highly relevant for determining that a banana is ripe. Iconic representations support the discrimination between sensory inputs, while categorical representations support the identification of categories (i.e., types) from category instances (i.e., tokens). Supporting categorical representations in LIDA may require the addition of weights to referential activation links and the development of a new learning rule for updating those weights.

### **ES-Hybrid's Predictions**

All theories—cognitive or otherwise—should make testable predictions. While I have no experience in experimental psychology, it seems reasonable to suppose that the phenomena of



eventual grounding may be testable (e.g., using neuroimaging). This may largely involve comparing the patterns of neural activation that occur when subjects first infer the existence of an unknown referent (e.g., when they are first presented a novel word form) to later patterns of activation that occur once grounding is established (e.g., when subjects are shown a depiction of that thing). The largest confounding factor would be the presence of non-referential (indexical) modal representations, which need to be separated from the referential (grounding) modal representations. Using separate sensory modalities for each may be one solution to this issue. Automatized internal behaviors (see Chapter 7) may be another testable prediction.

## References

- Adams, F., & Campbell, K. (1999). Modality and abstract concepts [Peer commentary on “Perceptual Symbol Systems,” by L. W. Barsalou]. *Behavioral and Brain Sciences*, 22(4), 610.
- Agrawal, P., Franklin, S., & Snider, J. (2018). Sensory memory for grounded representations in a cognitive architecture. In *Proceedings of the Sixth Annual Conference on Advances in Cognitive Systems (ACS Poster Collection)* (pp. 1–18).
- Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2623–2631.  
<https://doi.org/10.1145/3292500.3330701>
- Albright, T. D. (2012). On the perception of probable things: Neural substrates of associative memory, imagery, and perception. *Neuron*, 74(2), 227–245.
- Aleksander, I., De Gregorio, M., França, F. M. G., Lima, P. M. V., & Morton, H. (2009). A brief introduction to weightless neural systems. *ESANN*, 299–305.
- Anderson, J. R. (1978). Arguments concerning representations for mental imagery. *Psychological Review*, 85(4), 249–277. <https://doi.org/10.1037/0033-295X.85.4.249>
- Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, 111(4), 1036–1060.  
<https://doi.org/10.1037/0033-295X.111.4.1036>

- Angelini, M., Calbi, M., Ferrari, A., Sbriscia-Fioretti, B., Franca, M., Gallese, V., & Umiltà, M. A. (2015). Motor inhibition during overt and covert actions: An electrical neuroimaging study. *PloS One*, *10*(5), Article e0126800. <https://doi.org/10.1371/journal.pone.0126800>
- Baars, B. J. (1988). *A cognitive theory of consciousness*. Cambridge University Press.
- Baddeley, A. (1992). Working memory. *Science*, *255*(5044), 556–559.
- Baddeley, A., & Hitch, G. (1974). Working memory. In G. Bower (Ed.), *The psychology of learning and motivation* (Vol. 8, pp. 47–89). Academic Press.
- Baddeley, A., & Hitch, G. (1994). Developments in the concept of working memory. *Neuropsychology*, *8*(4), 485–493. <https://doi.org/10.1037/0894-4105.8.4.485>
- Bakker, M., De Lange, F. P., Stevens, J. A., Toni, I., & Bloem, B. R. (2007). Motor imagery of gait: A quantitative approach. *Experimental Brain Research*, *179*(3), 497–504.
- Bank, D., Koenigstein, N., & Giryes, R. (2021). *Autoencoders*. arXiv. <https://doi.org/10.48550/arXiv.2003.05991>
- Bar, M. (2009). The proactive brain: Memory for predictions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1521), 1235–1243.
- Baroni, M., & Lenci, A. (2010). Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, *36*(4), 673–721.
- Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, *22*(4), 577–660.
- Barsalou, L. W. (2003). Situated simulation in the human conceptual system. *Language and Cognitive Processes*, *18*(5–6), 513–562. <https://doi.org/10.1080/01690960344000026>
- Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology*, *59*(1), 617–645.

- Barsalou, L. W. (2009). Simulation, situated conceptualization, and prediction. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364, 1281–1289.
- Barsalou, L. W. (2010). Grounded cognition: Past, present, and future. *Topics in Cognitive Science*, 2(4), 716–724.
- Barsalou, L. W. (2016a). On staying grounded and avoiding quixotic dead ends. *Psychonomic Bulletin & Review*, 23(4), 1122–1142.
- Barsalou, L. W. (2016b). Situated conceptualization: Theory and applications. In *Foundations of embodied cognition. Volume 1, perceptual and emotional embodiment* (pp. 11–37). Psychology Press.
- Barsalou, L. W. (2020). Challenges and opportunities for grounding cognition. *Journal of Cognition*, 3(1), Article 31. <https://doi.org/10.5334/joc.116>
- Barsalou, L. W., Dutriaux, L., & Scheepers, C. (2018). Moving beyond the distinction between concrete and abstract concepts. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 373, Article 20170144. <https://doi.org/10.1098/rstb.2017.0144>
- Barsalou, L. W., Santos, A., Simmons, W. K., & Wilson, C. D. (2008). Language and simulation in conceptual processing. In *Symbols, embodiment, and meaning* (pp. 245–283). Oxford University Press.
- Barsalou, L. W., Simmons, W. K., Barbey, A. K., & Wilson, C. D. (2003). Grounding conceptual knowledge in modality-specific systems. *Trends in Cognitive Sciences*, 7(2), 84–91.
- Bartlett, F. (1958). *Thinking: An experimental and social study*. Basic Books.

- Beauchamp, M. S., Petit, L., Ellmore, T. M., Ingeholm, J., & Haxby, J. V. (2001). A parametric fMRI study of overt and covert shifts of visuospatial attention. *NeuroImage*, *14*(2), 310–321. <https://doi.org/10.1006/nimg.2001.0788>
- Bengio, Y., Courville, A. C., & Vincent, P. (2012). Unsupervised feature learning and deep learning: A review and new perspectives. *CoRR, Abs/1206.5538*, 1(2665), 2012.
- Bergen, B. (2012). *Louder than words: The new science of how the mind makes meaning*. Basic Books (AZ).
- Bergen, B. (2015). Embodiment, simulation and meaning. In *The Routledge handbook of semantics* (pp. 142–157).
- Bergen, B., & Chang, N. (2005). Embodied construction grammar in simulation-based language understanding. In J. O. Östman & M. Fried (Eds.), *Construction grammars: Cognitive grounding and theoretical extensions*. John Benjamins.
- Bergen, B., Lindsay, S., Matlock, T., & Narayanan, S. (2007). Spatial and linguistic aspects of visual imagery in sentence comprehension. *Cognitive Science*, *31*(5), Article 5.
- Bergen, B., & Wheeler, K. (2010). Grammatical aspect and mental simulation. *Brain and Language*, *112*(3), 150–158.
- Binder, J. R. (2016). In defense of abstract conceptual representations. *Psychonomic Bulletin & Review*, *23*(4), 1096–1108.
- Binder, J. R., Westbury, C. F., McKiernan, K. A., Possing, E. T., & Medler, D. A. (2005). Distinct brain systems for processing concrete and abstract concepts. *Journal of Cognitive Neuroscience*, *17*(6), 905–917.
- Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, *18*(2), 227–247.

- Borghini, A. M., Binkofski, F., Castelfranchi, C., Cimatti, F., Scorolli, C., & Tummolini, L. (2017). The challenge of abstract concepts. *Psychological Bulletin*, *143*, 263–292. <https://doi.org/10.1037/bul0000089>
- Borst, G., Kosslyn, S. M., & Denis, M. (2006). Different cognitive processes in two image-scanning paradigms. *Memory & Cognition*, *34*(3), 475–490. <https://doi.org/10.3758/BF03193572>
- Boureau, Y.-L., Ponce, J., & LeCun, Y. (2010). A theoretical analysis of feature pooling in visual recognition. *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 111–118.
- Braitenberg, V. (1986). *Vehicles: Experiments in synthetic psychology*. MIT press.
- Bratman, M. (1987). *Intention, plans, and practical reason* (Vol. 10). Harvard University Press Cambridge, MA.
- Brentano, F. (2012). *Psychology from an empirical standpoint*. Routledge. (Original work published 1874)
- Brewin, C. R., Gregory, J. D., Lipton, M., & Burgess, N. (2010). Intrusive images in psychological disorders: Characteristics, neural mechanisms, and treatment implications. *Psychological Review*, *117*, 210–232. <https://doi.org/10.1037/a0018113>
- Brooks, R. A. (1986). A robust layered control system for a mobile robot. *IEEE Journal on Robotics and Automation*, *2*, 14–23.
- Brooks, R. A. (1990). Elephants don't play chess. *Robotics and Autonomous Systems*, *6*, 3–15.
- Brooks, R. A. (1991a). How to build complete creatures rather than isolated cognitive simulators. In K. VanLehn (Ed.), *Architectures for intelligence: The 22nd Carnegie Mellon Symposium on Cognition* (pp. 225–240). Erlbaum.

- Brooks, R. A. (1991b). Intelligence without representation. *Artificial Intelligence*, 47(1–3), 139–159.
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. *ArXiv*. <http://arxiv.org/abs/2005.14165>
- Bruner, J. S., Postman, L., & Rodrigues, J. (1951). Expectation and the perception of color. *The American Journal of Psychology*, 64(2), 216–227.
- Buccino, G., Binkofski, F., Fink, G. R., Fadiga, L., Fogassi, L., Gallese, V., Seitz, R. J., Zilles, K., Rizzolatti, G., & Freund, H.-J. (2013). Action observation activates premotor and parietal areas in a somatotopic manner: An fMRI study. In *Social Neuroscience* (pp. 133–141). Psychology Press.
- Busse, L., Katzner, S., & Treue, S. (2008). Temporal dynamics of neuronal modulation during exogenous and endogenous shifts of visual attention in macaque area MT. *Proceedings of the National Academy of Sciences*, 105(42), 16380–16385.  
<https://doi.org/10.1073/pnas.0707369105>
- Caeyenberghs, K., Tsoupas, J., Wilson, P. H., & Smits-Engelsman, B. C. (2009). Motor imagery development in primary school children. *Developmental Neuropsychology*, 34(1), 103–121.
- Calvo-Merino, B., Glaser, D. E., Grèzes, J., Passingham, R. E., & Haggard, P. (2005). Action observation and acquired motor skills: An FMRI study with expert dancers. *Cerebral Cortex*, 15(8), 1243–1249.

- Calvo-Merino, B., Grèzes, J., Glaser, D. E., Passingham, R. E., & Haggard, P. (2006). Seeing or doing? Influence of visual and motor familiarity in action observation. *Current Biology*, *16*(19), 1905–1910.
- Cangelosi, A., & Riga, T. (2006). An embodied model for sensorimotor grounding and grounding transfer: Experiments with epigenetic robots. *Cognitive Science*, *30*(4), 673–689.
- Carrasco, M. (2011). Visual attention: The past 25 years. *Vision Research*, *51*(13), 1484–1525. <https://doi.org/10.1016/j.visres.2011.04.012>
- Cave, K. R., & Kosslyn, S. M. (1989). Varieties of size-specific visual selection. *Journal of Experimental Psychology: General*, *118*, 148–164. <https://doi.org/10.1037/0096-3445.118.2.148>
- Chandler, D. (2022). *Semiotics: The basics* (4th ed.). Routledge.
- Chella, A., Frixione, M., & Gaglio, S. (1997). A cognitive architecture for artificial vision. *Artificial Intelligence*, *89*(1), 73–111. [https://doi.org/10.1016/S0004-3702\(96\)00039-2](https://doi.org/10.1016/S0004-3702(96)00039-2)
- Chemero, A. (2013). Radical embodied cognitive science. *Review of General Psychology*, *17*(2), 145–150.
- Chen, W., Kato, T., Zhu, X.-H., Ogawa, S., Tank, D. W., & Ugurbil, K. (1998). Human primary visual cortex and lateral geniculate nucleus activation during visual imagery. *Neuroreport*, *9*(16), 3669–3674.
- Chica, A. B., Bartolomeo, P., & Lupiáñez, J. (2013). Two cognitive and neural systems for endogenous and exogenous spatial attention. *Behavioural Brain Research*, *237*, 107–123. <https://doi.org/10.1016/j.bbr.2012.09.027>



- Chun, M. M., & Jiang, Y. (1998). Contextual cueing: Implicit learning and memory of visual context guides spatial attention. *Cognitive Psychology*, *36*(1), 28–71.  
<https://doi.org/10.1006/cogp.1998.0681>
- Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 181–204.
- Clement, J. (1994). Use of physical intuition and imagistic simulation in expert problem solving. In D. Tirosh (Ed.), *Implicit and explicit knowledge* (pp. 204–244). Ablex.
- Cohen, M. S., Kosslyn, S. M., Breiter, H. C., DiGirolamo, G. J., Thompson, W. L., Anderson, A. K., Bookheimer, S. Y., Rosen, B. R., & Belliveau, J. W. (1996). Changes in cortical activity during mental rotation: A mapping study using functional MRI. *Brain*, *119*(1), 89–100. <https://doi.org/10.1093/brain/119.1.89>
- Cooper, L. A. (1975). Mental rotation of random two-dimensional shapes. *Cognitive Psychology*, *7*(1), 20–43.
- Cooper, L. A. (1976). Demonstration of a mental analog of an external rotation. *Perception & Psychophysics*, *19*(4), 296–302.
- Cope, T. E., Sohoglu, E., Sedley, W., Patterson, K., Jones, P. S., Wiggins, J., Dawson, C., Grube, M., Carlyon, R. P., & Griffiths, T. D. (2017). Evidence for causal top-down frontal contributions to predictive processes in speech perception. *Nature Communications*, *8*(1), 1–16.
- Coulom, R. (2006). Efficient selectivity and backup operators in Monte-Carlo tree search. *Computers and Games: 5th International Conference*, 72–83.
- Crapse, T. B., & Sommer, M. A. (2008). Corollary discharge across the animal kingdom. *Nature Reviews Neuroscience*, *9*(8), 587–600.

- Craver-Lemley, C., & Reeves, A. (1992). How visual imagery interferes with vision. *Psychological Review*, 99(4), 633–649. <https://doi.org/10.1037/0033-295X.99.4.633>
- Crutch, S. J., & Warrington, E. K. (2005). Abstract and concrete concepts have structurally different representational frameworks. *Brain*, 128(3), 615–627.
- Cutsuridis, V., Hussain, A., & Taylor, J. G. (2011). *Perception-action cycle: Models, architectures, and hardware*. Springer Science & Business Media.
- Cybenko, G. (1988). *Continuous valued neural networks with two hidden layers are sufficient* [Technical Report]. Tufts University.
- Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4), 303–314.
- Cyc's knowledge base – Cycorp inc. (n.d.). Retrieved December 15, 2020, from <https://www.cyc.com/archives/service/cyc-knowledge-base>
- Damasio, A. R. (1989). Time-locked multiregional retroactivation: A systems-level proposal for the neural substrates of recall and recognition. *Cognition*, 33(1–2), 25–62.
- de Haan, B., Morgan, P. S., & Rorden, C. (2008). Covert orienting of attention and overt eye movements activate identical brain regions. *Brain Research*, 1204, 102–111. <https://doi.org/10.1016/j.brainres.2008.01.105>
- De Vega, M., Glenberg, A., & Graesser, A. (2012). *Symbols and embodiment: Debates on meaning and cognition*. Oxford University Press.
- Deacon, T. W. (1997). *The symbolic species: The co-evolution of language and the brain*. WW Norton & Company.
- Decety, J., Jeannerod, M., & Prablanc, C. (1989). The timing of mentally represented actions. *Behavioural Brain Research*, 34(1–2), 35–42.

- Decety, J., Perani, D., Jeannerod, M., Bettinardi, V., Tadary, B., Woods, R., Mazziotta, J. C., & Fazio, F. (1994). Mapping motor representations with positron emission tomography. *Nature*, *371*(6498), 600–602.
- Desai, R. H., Reilly, M., & van Dam, W. (2018). The multifaceted abstract brain. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1752), 20170122.
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*.  
<https://arxiv.org/abs/1810.04805>
- Dijkstra, N., Zeidman, P., Ondobaka, S., van Gerven, M. A., & Friston, K. (2017). Distinct top-down and bottom-up brain connectivity during visual perception and imagery. *Scientific Reports*, *7*(1), 1–9.
- Dijkstra, T. M. H., Schöner, G., & Gielen, C. (1994). Temporal stability of the action-perception cycle for postural control in a moving visual environment. *Experimental Brain Research*, *97*(3), 477–486.
- D’Mello, S. K., Ramamurthy, U., Negatu, A., & Franklin, S. (2006). A procedural learning mechanism for novel skill acquisition. *Proceeding of Adaptation in Artificial and Biological Systems, Aisb*, *6*, 184–185.
- Dong, D., & Franklin, S. (2015). A new action execution module for the learning intelligent distribution agent (LIDA): The sensory motor system. *Cognitive Computation*, *7*(5), 552–568.
- Dong, D., Franklin, S., & Agrawal, P. (2015). Estimating human movements using memory of errors. *Procedia Computer Science*, *71*, 1–10.

- Dove, G. (2009). Beyond perceptual symbols: A call for representational pluralism. *Cognition*, 110(3), 412–431.
- Dove, G. (2011). On the need for embodied and dis-embodied cognition. *Frontiers in Psychology*, 1, Article 242. <https://doi.org/10.3389/fpsyg.2010.00242>
- Dove, G. (2016). Three symbol ungrounding problems: Abstract concepts and the future of embodied cognition. *Psychonomic Bulletin & Review*, 23(4), 1109–1121.
- Drescher, G. L. (1991). *Made-up minds: A constructivist approach to artificial intelligence*. MIT press.
- Dreyfus, H. L. (2002). Intelligence without representation—Merleau-Ponty’s critique of mental representation: The relevance of phenomenology to scientific explanation. *Phenomenology and the Cognitive Sciences*, 1, 367–383.
- Driskell, J. E., Copper, C., & Moran, A. (1994). Does mental practice enhance performance? *Journal of Applied Psychology*, 79(4), 481–492. <https://doi.org/10.1037/0021-9010.79.4.481>
- Edelman, G. M. (1987). *Neural Darwinism: The theory of neuronal group selection*. Basic books.
- Ehrsson, H. H., Geyer, S., & Naito, E. (2003). Imagery of voluntary movement of fingers, toes, and tongue activates corresponding body-part-specific motor representations. *Journal of Neurophysiology*, 90(5), 3304–3316. <https://doi.org/10.1152/jn.01113.2002>
- Eng, H. Y., Chen, D., & Jiang, Y. (2005). Visual working memory for simple and complex visual stimuli. *Psychonomic Bulletin & Review*, 12, 1127–1133.

- Eriksen, C. W., & Murphy, T. D. (1987). Movement of attentional focus across the visual field: A critical look at the evidence. *Perception & Psychophysics*, *42*, 299–305.  
<https://doi.org/10.3758/BF03203082>
- Eriksen, C. W., & St. James, J. D. (1986). Visual attention within and around the field of focal attention: A zoom lens model. *Perception & Psychophysics*, *40*(4), 225–240.  
<https://doi.org/10.3758/BF03211502>
- Erk, K. (2012). Vector space models of word meaning and phrase meaning: A survey. *Language and Linguistics Compass*, *6*(10), 635–653.
- Ernst, G. W., & Newell, A. (1969). *GPS: A case study in generality and problem solving*. Academic Press.
- Eslami, S. M. A., Jimenez Rezende, D., Besse, F., Viola, F., Morcos, A. S., Garnelo, M., Ruderman, A., Rusu, A. A., Danihelka, I., Gregor, K., Reichert, D. P., Buesing, L., Weber, T., Vinyals, O., Rosenbaum, D., Rabinowitz, N., King, H., Hillier, C., Botvinick, M., ... Hassabis, D. (2018). Neural scene representation and rendering. *Science*, *360*(6394), 1204–1210. <https://doi.org/10.1126/science.aar6170>
- Farah, M. J. (1985). Psychophysical evidence for a shared representational medium for mental images and percepts. *Journal of Experimental Psychology: General*, *114*(1), 91–103.  
<https://doi.org/10.1037/0096-3445.114.1.91>
- Farah, M. J. (1989). Mechanisms of imagery-perception interaction. *Journal of Experimental Psychology: Human Perception and Performance*, *15*(2), 203–211.  
<https://doi.org/10.1037/0096-1523.15.2.203>

- Finke, R. A., & Pinker, S. (1982). Spontaneous imagery scanning in mental extrapolation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 8(2), 142–147.  
<https://doi.org/10.1037/0278-7393.8.2.142>
- Finke, R. A., & Pinker, S. (1983). Directional scanning of remembered visual patterns. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 9(3), 398–410.  
<https://doi.org/10.1037/0278-7393.9.3.398>
- Fisher, J. C. (2006). Does simulation theory really involve simulation? *Philosophical Psychology*, 19(4), 417–432.
- Fitts, P. M., & Posner, M. I. (1967). *Human performance*. Brooks/Cole.
- Fodor, J. A. (1975). *The language of thought*. Thomas Y. Crowell Company.
- Fodor, J. A. (2008). *LOT 2: The language of thought revisited*. Oxford University Press.
- Fougnie, D., Suchow, J. W., & Alvarez, G. A. (2012). Variability in the quality of visual working memory. *Nature Communications*, 3(1), Article 1.  
<https://doi.org/10.1038/ncomms2237>
- Frak, V., Paulignan, Y., & Jeannerod, M. (2001). Orientation of the opposition axis in mentally simulated grasping. *Experimental Brain Research*, 136(1), 120–127.
- Franklin, S. (1995). *Artificial Minds* (1st ed.). MIT Press.
- Franklin, S. (2003). IDA, a conscious artifact? *Journal of Consciousness Studies*, 10(4–5), 47–66.
- Franklin, S., & Baars, B. (2010). Two varieties of unconscious processes. In E. Perry, D. Collerton, H. Ashton, & F. LeBeau (Eds.), *New horizons in the neuroscience of consciousness* (pp. 91–102). John Benjamins.

- Franklin, S., & Graesser, A. (1997). Is it an agent, or just a program?: A taxonomy for autonomous agents. In *Proceedings of the Third International Workshop on Agent Theories, Architectures, and Languages* (pp. 21–35). Springer-Verlag.
- Franklin, S., Madl, T., D’mello, S., & Snider, J. (2013). LIDA: A systems-level architecture for cognition, emotion, and learning. *IEEE Transactions on Autonomous Mental Development*, 6(1), 19–41.
- Franklin, S., Madl, T., Strain, S., Faghihi, U., Dong, D., Kugele, S., Snider, J., Agrawal, P., & Chen, S. (2016). A LIDA cognitive model tutorial. *Biologically Inspired Cognitive Architectures*, 16, 105–130.
- Franklin, S., Strain, S., McCall, R., & Baars, B. (2013). Conceptual commitments of the LIDA model of cognition. *Journal of Artificial General Intelligence*, 4(2), 1–22.
- Freeman, W. J. (2002). The limbic action-perception cycle controlling goal-directed animal behavior. *Proceedings of the 2002 International Joint Conference on Neural Networks*, 3, 2249–2254. <https://doi.org/10.1109/IJCNN.2002.1007491>
- Frick, A., Daum, M. M., Wilson, M., & Wilkening, F. (2009). Effects of action on children’s and adults’ mental imagery. *Journal of Experimental Child Psychology*, 104(1), 34–51. <https://doi.org/10.1016/j.jecp.2009.01.003>
- Fujita, I., Tanaka, K., Ito, M., & Cheng, K. (1992). Columns for visual features of objects in monkey inferotemporal cortex. *Nature*, 360(6402), 343–346.
- Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36, 193–202.

- Fuster, J. M. (2004). Upper processing stages of the perception–action cycle. *Trends in Cognitive Sciences*, 8, 143–145.
- Gallagher, S. (2008). Are minimal representations still representations? *International Journal of Philosophical Studies*, 16(3), 351–369.
- Gallagher, S. (2015). Invasion of the body snatchers: How embodied cognition is being disembodied. *The Philosophers' Magazine*, 68, 96–102.
- Gallagher, S. (2017). *Enactivist interventions: Rethinking the mind*. Oxford University Press.
- Gallese, V. (2003). The roots of empathy: The shared manifold hypothesis and the neural basis of intersubjectivity. *Psychopathology*, 36(4), 171–180.
- Gallese, V. (2005). Embodied simulation: From neurons to phenomenal experience. *Phenomenology and the Cognitive Sciences*, 4(1), 23–48.
- Gallese, V. (2007). Before and below ‘theory of mind’: Embodied simulation and the neural correlates of social cognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362(1480), 659–669. <https://doi.org/10.1098/rstb.2006.2002>
- Gallese, V. (2018). Embodied simulation and its role in cognition. *Reti, saperi, linguaggi, Italian Journal of Cognitive Sciences*, 1/2018, 31–46. <https://doi.org/10.12832/90969>
- Gallese, V., Fadiga, L., Fogassi, L., & Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain*, 119(2), 593–609.
- Ganis, G., Keenan, J. P., Kosslyn, S. M., & Pascual-Leone, A. (2000). Transcranial magnetic stimulation of primary motor cortex affects mental rotation. *Cerebral Cortex*, 10(2), 175–180.



- Ganis, G., Thompson, W. L., & Kosslyn, S. M. (2004). Brain areas underlying visual mental imagery and visual perception: An fMRI study. *Cognitive Brain Research*, 20(2), 226–241.
- Garcez, A. d'Avila, & Lamb, L. C. (2020). *Neurosymbolic AI: The 3rd wave*. arXiv.  
<https://doi.org/10.48550/arXiv.2012.05876>
- Georgopoulos, A. P., Lurito, J. T., Petrides, M., Schwartz, A. B., & Massey, J. T. (1989). Mental rotation of the neuronal population vector. *Science*, 243(4888), 234–236.
- Goertzel, B., & Pennachin, C. (2007). The Novamente artificial intelligence engine. In *Artificial general intelligence* (pp. 63–129). Springer, Berlin, Heidelberg.
- Goldman, A. I. (1992). In defense of the simulation theory. *Mind & Language*, 7(1–2), 104–119.  
<https://doi.org/10.1111/j.1468-0017.1992.tb00200.x>
- Goldman, A. I. (2012). A moderate approach to embodied cognitive science. *Review of Philosophy and Psychology*, 3(1), 71–88.
- Goldman, A. I. (2013). The bodily formats approach to embodied cognition. In *Current controversies in philosophy of mind* (pp. 91–108). Routledge.
- Goldstone, R. L., & Barsalou, L. W. (1998). Reuniting perception and conception. *Cognition*, 65(2–3), 231–262.
- Goodale, M. A., & Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences*, 15(1), 20–25.
- Goodfellow, I., Mirza, M., Xiao, D., Courville, A., & Bengio, Y. (2015). *An empirical investigation of catastrophic forgetting in gradient-based neural networks*. arXiv.  
<https://doi.org/10.48550/arXiv.1312.6211>

- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, 2672–2680.
- Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing? *International Journal of Human-Computer Studies*, 43(5), 907–928.  
<https://doi.org/10.1006/ijhc.1995.1081>
- Grush, R. (2004). The emulation theory of representation: Motor control, imagery, and perception. *Behavioral and Brain Sciences*, 27(3), 377–396.  
<https://doi.org/10.1017/S0140525X04000093>
- Haimovici, S. (2018). The modal-amodal distinction in the debate on conceptual format. *Philosophies*, 3(2), Article 7. <https://doi.org/10.3390/philosophies3020007>
- Hanin, B. (2019). Universal function approximation by deep neural nets with bounded width and ReLU activations. *Mathematics*, 7(10), Article 992. <https://doi.org/10.3390/math7100992>
- Hansen, T., Olkkonen, M., Walter, S., & Gegenfurtner, K. R. (2006). Memory modulates color appearance. *Nature Neuroscience*, 9(11), 1367–1368.
- Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42, 335–346.
- Harris, Z. S. (1954). Distributional structure. *Word*, 10(2–3), 146–162.
- Hart, P. E., Nilsson, N. J., & Raphael, B. (1968). A formal basis for the heuristic determination of minimum cost paths. *IEEE Transactions on Systems Science and Cybernetics*, 4(2), 100–107.
- Hawkins, J., Ahmad, S., & Dubinsky, D. (2010). *Hierarchical temporal memory including HTM cortical learning algorithms*. Numenta, Inc.

- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.
- Hesslow, G. (2002). Conscious thought as simulation of behaviour and perception. *Trends in Cognitive Sciences*, 6(6), 242–247. [https://doi.org/10.1016/S1364-6613\(02\)01913-7](https://doi.org/10.1016/S1364-6613(02)01913-7)
- Hesslow, G. (2012). The current status of the simulation theory of cognition. *Brain Research*, 1428, 71–79.
- Higgins, I., Amos, D., Pfau, D., Racaniere, S., Matthey, L., Rezende, D., & Lerchner, A. (2018). *Towards a definition of disentangled representations*. arXiv. <https://doi.org/10.48550/arXiv.1812.02230>
- Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Sharik, M., & Lerchner, A. (2017). Beta-VAE: Learning basic visual concepts with a constrained variational framework. *ICLR*, 2, 6.
- Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. R. (2012). *Improving neural networks by preventing co-adaptation of feature detectors*. arXiv. <https://doi.org/10.48550/arXiv.1207.0580>
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780.
- Hoffman, J. E., & Subramaniam, B. (1995). The role of visual attention in saccadic eye movements. *Perception & Psychophysics*, 57(6), 787–795. <https://doi.org/10.3758/BF03206794>
- Hofstadter, D. R., & Mitchell, M. (1994). The Copycat project: A model of mental fluidity and analogy-making. In *Analogical connections* (pp. 31–112). Ablex Publishing.

- Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2), 251–257.
- Iacoboni, M., Molnar-Szakacs, I., Gallese, V., Buccino, G., Mazziotta, J. C., & Rizzolatti, G. (2005). Grasping the intentions of others with one's own mirror neuron system. *PLOS Biology*, 3(3), Article e79. <https://doi.org/10.1371/journal.pbio.0030079>
- Intaitė, M., Noreika, V., Šoliūnas, A., & Falter, C. M. (2013). Interaction of bottom-up and top-down processes in the perception of ambiguous figures. *Vision Research*, 89, 24–31.
- Jacob, P. (2020). Intentionality. In *The Stanford Encyclopedia of Philosophy*. <https://plato.stanford.edu/archives/win2019/entries/intentionality/>
- Jeannerod, M. (1994). The representing brain: Neural correlates of motor intention and imagery. *Behavioral and Brain Sciences*, 17(2), 187–202.
- Jeannerod, M. (1995). Mental imagery in the motor context. *Neuropsychologia*, 33(11), 1419–1432.
- Jeannerod, M. (2001). Neural simulation of action: A unifying mechanism for motor cognition. *Neuroimage*, 14(1), S103–S109.
- Jeannerod, M. (2006). *Motor cognition: What actions tell the self*. Oxford University Press.
- Jeannerod, M., & Frak, V. (1999). Mental imaging of motor activity in humans. *Current Opinion in Neurobiology*, 9(6), 735–739.
- Jeannerod, M., Kennedy, H., & Magnin, M. (1979). Corollary discharge: Its possible implications in visual and oculomotor interactions. *Neuropsychologia*, 17(2), 241–258.
- Jirak, D., Menz, M. M., Buccino, G., Borghi, A. M., & Binkofski, F. (2010). Grasping language—a short story on embodiment. *Consciousness and Cognition*, 19(3), 711–720.
- Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

- Kautz, H. (2022). The third AI summer: AAAI Robert S. Engelmore Memorial Lecture. *AI Magazine*, 43(1), 93–104.
- Keller, P. E. (2012). Mental imagery in music performance: Underlying mechanisms and potential benefits. *Annals of the New York Academy of Sciences*, 1252(1), 206–213.
- Kingma, D. P., & Welling, M. (2013). *Auto-encoding variational Bayes*. arXiv.  
<https://doi.org/10.48550/arXiv.1312.6114>
- Kirkpatrick, S., Gelatt Jr, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, 220(4598), 671–680.
- Kirsh, D., & Maglio, P. (1994). On distinguishing epistemic from pragmatic action. *Cognitive Science*, 18, 513–549.
- Klein, I., Dubois, J., Mangin, J.-F., Kherif, F., Flandin, G., Poline, J.-B., Denis, M., Kosslyn, S. M., & Le Bihan, D. (2004). Retinotopic organization of visual mental images as revealed by functional magnetic resonance imaging. *Cognitive Brain Research*, 22(1), 26–31.
- Kosslyn, S. M. (1980). *Image and mind*. Harvard University Press.
- Kosslyn, S. M. (1994). *Image and brain: The resolution of the imagery debate* (1st ed.). MIT Press.
- Kosslyn, S. M., Ball, T. M., & Reiser, B. J. (1978). Visual images preserve metric spatial information: Evidence from studies of image scanning. *Journal of Experimental Psychology: Human Perception and Performance*, 4(1), 47.
- Kosslyn, S. M., Cave, C. B., Provost, D. A., & von Gierke, S. M. (1988). Sequential processes in image generation. *Cognitive Psychology*, 20(3), 319–343.
- Kosslyn, S. M., Margolis, J. A., Barrett, A. M., Goldknopf, E. J., & Daly, P. F. (1990). Age differences in imagery abilities. *Child Development*, 61(4), 995–1010.

- Kosslyn, S. M., Pascual-Leone, A., Felician, O., Camposano, S., Keenan, J. P., Ganis, G., Sukel, K. E., & Alpert, N. M. (1999). The role of area 17 in visual imagery: Convergent evidence from PET and rTMS. *Science*, *284*(5411), 167–170.
- Kosslyn, S. M., Reiser, B. J., Farah, M. J., & Fliegel, S. L. (1983). Generating visual images: Units and relations. *Journal of Experimental Psychology: General*, *112*, 278–303.  
<https://doi.org/10.1037/0096-3445.112.2.278>
- Kosslyn, S. M., Thompson, W. L., & Alpert, N. M. (1997). Neural systems shared by visual imagery and visual perception: A positron emission tomography study. *Neuroimage*, *6*(4), 320–334.
- Kosslyn, S. M., Thompson, W. L., & Ganis, G. (2006). *The case for mental imagery*. Oxford University Press.
- Kosslyn, S. M., Thompson, W. L., Klm, I. J., & Alpert, N. M. (1995). Topographical representations of mental images in primary visual cortex. *Nature*, *378*(6556), 496–498.
- Kosslyn, S. M., Thompson, W. L., Wraga, M., & Alpert, N. M. (2001). Imagining rotation by endogenous versus exogenous forces: Distinct neural mechanisms. *NeuroReport*, *12*(11), 2519–2525.
- Kosslyn, S. M., Thompson, W., Shephard, J., Ganis, G., Bell, D., Danovitch, J., Wittenberg, L., & Alpert, N. (2004). Brain rCBF and performance in visual imagery tasks: Common and distinct processes. *European Journal of Cognitive Psychology*, *16*(5), 696–716.
- Kotseruba, I., & Tsotsos, J. K. (2018). 40 years of cognitive architectures: Core cognitive abilities and practical applications. *Artificial Intelligence Review*, 1–78.
- Kreiman, G., Koch, C., & Fried, I. (2000). Imagery neurons in the human brain. *Nature*, *408*, 357–361. <https://doi.org/10.1038/35042575>

- Kringelbach, M. L., & Berridge, K. C. (2009). Towards a functional neuroanatomy of pleasure and happiness. *Trends in Cognitive Sciences*, 13(11), 479–487.
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).
- Kugele, S., & Franklin, S. (2020a). A study in activation: Towards a common lexicon and functional taxonomy in cognitive architectures. *Proceedings of the 18th Annual Meeting of the International Conference on Cognitive Modelling*, 138–144. <https://iccm-conference.neocities.org/2020/ICCM2020Proceedings.pdf>
- Kugele, S., & Franklin, S. (2020b). “Conscious” multi-modal perceptual learning for grounded simulation-based cognition. In S. Denison, M. Mack, Y. Xu, & B. C. Armstrong (Eds.), *Proceedings of the 42nd Annual Conference of the Cognitive Science Society* (pp. 2459–2465). Cognitive Science Society.
- Kugele, S., & Franklin, S. (2021). Learning in LIDA. *Cognitive Systems Research*, 66, 176–200. <https://doi.org/10.1016/j.cogsys.2020.11.001>
- Laird, John. E. (2012). *The Soar cognitive architecture*. MIT press.
- Lakoff, G., & Johnson, M. (1999). *Philosophy in the flesh: The embodied mind and its challenge to western thought*. Basic books.
- Lakoff, G., & Johnson, M. (2008). *Metaphors we live by*. University of Chicago press. (Original work published 1980)
- Lallee, S., & Dominey, P. F. (2013). Multi-modal convergence maps: From body schema and self-representation to mental imagery. *Adaptive Behavior*, 21(4), 274–285.

- Landauer, T. K. (2007). LSA as a theory of meaning. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis*. Psychology Press.
- Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211–240. <https://doi.org/10.1037/0033-295X.104.2.211>
- Lathrop, S. D., & Laird, J. E. (2007). Towards incorporating visual imagery into a cognitive architecture. *Proceedings of the Eighth International Conference on Cognitive Modeling*. ICCM 2007, Oxford, UK.
- LeCun, Y., & Bengio, Y. (1995). Convolutional networks for images, speech, and time series. In M. A. Arbib (Ed.), *The handbook of brain theory and neural networks* (pp. 276–279). MIT Press.
- LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., & Jackel, L. (1989). Handwritten digit recognition with a back-propagation network. In D. Touretzky (Ed.), *Advances in Neural Information Processing Systems* (Vol. 2). Morgan-Kaufmann. [https://proceedings.neurips.cc/paper\\_files/paper/1989/file/53c3bce66e43be4f209556518c2fcb54-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/1989/file/53c3bce66e43be4f209556518c2fcb54-Paper.pdf)
- LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.
- Lee, D. N. (1980). The optic flow field: The foundation of vision. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, *290*(1038), 169–179.
- Lenci, A. (2008). Distributional semantics in linguistic and cognitive research. *Italian Journal of Linguistics*, *20*(1), 1–31.



- Leshno, M., Lin, V. Y., Pinkus, A., & Schocken, S. (1993). Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*, 6(6), 861–867.
- Lewin, K. (1951). *Field theory in social science: Selected theoretical papers*.
- Lotze, M., Montoya, P., Erb, M., Hülsmann, E., Flor, H., Klose, U., Birbaumer, N., & Grodd, W. (1999). Activation of cortical and cerebellar motor areas during executed and imagined hand movements: An fMRI study. *Journal of Cognitive Neuroscience*, 11(5), 491–501.
- Louwerse, M. M. (2011). Symbol interdependency in symbolic and embodied cognition. *Topics in Cognitive Science*, 3(2), 273–302.
- Louwerse, M. M. (2018). Knowing the meaning of a word by the linguistic and perceptual company it keeps. *Topics in Cognitive Science*, 10(3), 573–589.
- Louwerse, M. M., & Jeuniaux, P. (2008). Language comprehension is both embodied and symbolic. *Symbols and Embodiment: Debates on Meaning and Cognition*, 309–326.
- Lu, Z., Pu, H., Wang, F., Hu, Z., & Wang, L. (2017). The expressive power of neural networks: A view from the width. *Advances in Neural Information Processing Systems*, 6231–6239.
- Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, 390, 279–281. <https://doi.org/10.1038/36846>
- Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, 28(2), 203–208. <https://doi.org/10.3758/BF03204766>
- Machery, E. (2006). Two dogmas of neo-empiricism. *Philosophy Compass*, 1(4), 398–412.
- Machery, E. (2016). The amodal brain and the offloading hypothesis. *Psychonomic Bulletin & Review*, 23(4), 1090–1095.

- Madl, T., Baars, B. J., & Franklin, S. (2011). The timing of the cognitive cycle. *PLOS ONE*, 6(4), Article e14803. <https://doi.org/10.1371/journal.pone.0014803>
- Madl, T., Franklin, S., Chen, K., Montaldi, D., & Trapp, R. (2016). Towards real-world capable spatial memory in the LIDA cognitive architecture. *Biologically Inspired Cognitive Architectures*, 16, 87–104. <https://doi.org/10.1016/j.bica.2016.02.001>
- Madl, T., Franklin, S., Chen, K., & Trapp, R. (2018). A computational cognitive framework of spatial memory in brains and robots. *Cognitive Systems Research*, 47, 147–172.
- Maes, P. (1989). How to do the right thing. *Connection Science*, 1(3), 291–323.
- Maes, P. (1991). The agent network architecture (ANA). *Acm Sigart Bulletin*, 2(4), 115–120.
- Mahendran, A., & Vedaldi, A. (2016). Visualizing deep convolutional neural networks using natural pre-images. *International Journal of Computer Vision*, 120(3), 233–255.
- Mahon, B. Z., & Caramazza, A. (2008). A critical look at the embodied cognition hypothesis and a new proposal for grounding conceptual content. *Journal of Physiology-Paris*, 102(1–3), 59–70.
- Mahon, B. Z., & Hickok, G. (2016). Arguments about the nature of concepts: Symbols, embodiment, and beyond. *Psychonomic Bulletin & Review*, 23(4), 941–958.
- Mao, J., Gan, C., Kohli, P., Tenenbaum, J. B., & Wu, J. (2019). *The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision*. arXiv. <https://doi.org/10.48550/arXiv.1904.12584>
- Marcus, G. (2018). *Deep learning: A critical appraisal*. arXiv. <https://doi.org/10.48550/arXiv.1801.00631>
- Marcus, G. (2020). *The next decade in AI: Four steps towards robust artificial intelligence*. arXiv. <https://doi.org/10.48550/arXiv.2002.06177>

- Mast, F. W., Berthoz, A., & Kosslyn, S. M. (2001). Mental imagery of visual motion modifies the perception of rollvection stimulation. *Perception, 30*(8), 945–957.
- McCall, R., Franklin, S., Faghihi, U., Snaider, J., & Kugele, S. (2020). Artificial motivation for cognitive software agents. *Journal of Artificial General Intelligence, 11*(1), 38–69.
- McCall, R., Franklin, S., & Friedlander, D. (2010). Grounded event-based and modal representations for objects, relations, beliefs, etc. *Twenty-Third International FLAIRS Conference*.
- McCall, R., Snaider, J., & Franklin, S. (2010). *Sensory and perceptual scene representation* [Unpublished].
- McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation* (Vol. 24, pp. 109–165). Elsevier.
- Mechelli, A., Price, C. J., Friston, K. J., & Ishai, A. (2004). Where bottom-up meets top-down: Neuronal interactions during perception and imagery. *Cerebral Cortex, 14*(11), 1256–1265. <https://doi.org/10.1093/cercor/bhh087>
- Metzler, J., & Shepard, R. N. (1974). Transformational studies of the internal representation of three-dimensional objects. In *Theories in cognitive psychology: The Loyola Symposium*. Lawrence Erlbaum.
- Michel, C. (2021). Overcoming the modal/amodal dichotomy of concepts. *Phenomenology and the Cognitive Sciences, 20*, 655–677. <https://doi.org/10.1007/s11097-020-09678-y>
- Miller, G. A. (2003). The cognitive revolution: A historical perspective. *Trends in Cognitive Sciences, 7*(3), 141–144.
- Minsky, M. (1961). Steps toward artificial intelligence. *Proceedings of the IRE, 49*(1), 8–30.

- Minsky, M. (1975). A framework for representing knowledge. In *The Psychology of Computer Vision* (pp. 211–277). McGraw-Hill.
- Minsky, M. (1986). *The society of mind*. Simon and Schuster.
- Mishkin, M., Ungerleider, L. G., & Macko, K. A. (1983). Object vision and spatial vision: Two cortical pathways. *Trends in Neurosciences*, 6, 414–417.
- Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., & Ostrovski, G. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540), 529–533.
- Molina, M., Tijus, C., & Jouen, F. (2008). The emergence of motor imagery in children. *Journal of Experimental Child Psychology*, 99(3), 196–209.
- Mondor, T. A., & Zatorre, R. J. (1995). Shifting and focusing auditory spatial attention. *Journal of Experimental Psychology: Human Perception and Performance*, 21(2), 387.
- Moulton, S. T., & Kosslyn, S. M. (2009). Imagining predictions: Mental imagery as mental emulation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1521), 1273–1280.
- Murphy, G. (2004). *The big book of concepts*. MIT press.
- Naimi, A. I., & Balzer, L. B. (2018). Stacked generalization: An introduction to super learning. *European Journal of Epidemiology*, 33(5), 459–464.
- Neemeh, Z. A., Kronsted, C., Kugele, S., & Franklin, S. (2021). Body schema in autonomous agents. *Journal of Artificial Intelligence and Consciousness*, 8(1), 113–145.  
<https://doi.org/10.1142/S2705078521500065>
- Negatu, A., & Franklin, S. (2002). An action selection mechanism for 'conscious' software agents. *Cognitive Science Quarterly*, 2, 363–386.

- Neisser, U. (1976). *Cognition and reality: Principles and implications of cognitive psychology*. W. H. Freeman.
- Newell, A. (1973). You can't play 20 questions with nature and win: Projective comments on the papers of this symposium. In *Visual information processing*. Academic Press.
- Newell, A. (1994). *Unified theories of cognition*. Harvard University Press.
- Newell, A., Shaw, J. C., & Simon, H. A. (1957). *Empirical explorations of the logic theory machine: A case study*. 218–230.
- Newell, A., & Simon, H. A. (1956). The logic theory machine: A complex information processing system. *Institute of Radio Engineers Transactions on Information Theory*, 2, 61–79.
- Newell, A., & Simon, H. A. (1961). Computer simulation of human thinking: A theory of problem solving expressed as a computer program permits simulation of thinking processes. *Science*, 134(3495), 2011–2017.
- Newell, A., & Simon, H. A. (1976). Computer science as empirical inquiry: Symbols and search. *Communication of the ACM*, 19, 113–126.
- Newen, A., De Bruin, L., & Gallagher, S. (2018). *The Oxford handbook of 4E cognition*. Oxford University Press.
- Nobre, A. C., Gitelman, D. R., Dias, E. C., & Mesulam, M. M. (2000). Covert visual spatial orienting and saccades: Overlapping neural systems. *NeuroImage*, 11(3), 210–216. <https://doi.org/10.1006/nimg.2000.0539>
- Norton, A. (1995). Dynamics: An introduction. In R. F. Port & T. Van Gelder (Eds.), *Mind as motion: Explorations in the dynamics of cognition* (Vol. 1, pp. 45–68). MIT Press.

- O’Callaghan, C., Kveraga, K., Shine, J. M., Adams Jr, R. B., & Bar, M. (2017). Predictions penetrate perception: Converging insights from brain, behaviour and disorder. *Consciousness and Cognition*, *47*, 63–74.
- Ohlsson, S. (1999). Selecting is not abstracting [Peer commentary on “Perceptual Symbol Systems,” by L. W. Barsalou]. *Behavioral and Brain Sciences*, *22*(4), 630–631.
- Oord, A. van den, Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., & Kavukcuoglu, K. (2016). *WaveNet: A generative model for raw audio*. arXiv. <https://doi.org/10.48550/arXiv.1609.03499>
- Paivio, A. (1990). *Mental representations: A dual coding approach*. Oxford University Press. (Original work published 1986)
- Paivio, A. (2014). Intelligence, dual coding theory, and the brain. *Intelligence*, *47*, 141–158.
- Passingham, R. E. (1988). Premotor cortex and preparation for movement. *Experimental Brain Research*, *70*(3), 590–596. <https://doi.org/10.1007/BF00247607>
- Patterson, K., Nestor, P. J., & Rogers, T. T. (2007). Where do you know what you know? The representation of semantic knowledge in the human brain. *Nature Reviews Neuroscience*, *8*(12), 976–987.
- Pearson, J., Clifford, C. W. G., & Tong, F. (2008). The functional impact of mental imagery on conscious perception. *Current Biology*, *18*(13), 982–986. <https://doi.org/10.1016/j.cub.2008.05.048>
- Peirce, C. S. (1992). *The essential Peirce: Selected philosophical writings* (N. Houser & C. Kloesel, Eds.; Vol. 1). Indiana University Press. (Original work published 1867–1893)
- Peirce, C. S. (1998). *The essential Peirce: Selected philosophical writings* (the Peirce Edition Project, Ed.; Vol. 2). Indiana University Press. (Original work published 1893–1913)

- Perky, C. W. (1910). An experimental study of imagination. *The American Journal of Psychology*, 21(3), 422–452.
- Pezzulo, G., Barsalou, L. W., Cangelosi, A., Fischer, M. H., McRae, K., & Spivey, M. (2013). Computational grounded cognition: A new alliance between grounded cognition and computational modeling. *Frontiers in Psychology*, 3, Article 612.
- Pezzulo, G., & Castelfranchi, C. (2007). The symbol detachment problem. *Cognitive Processing*, 8(2), 115–131. <https://doi.org/10.1007/s10339-007-0164-0>
- Pfurtscheller, G., & Neuper, C. (1997). Motor imagery activates primary sensorimotor area in humans. *Neuroscience Letters*, 239(2–3), 65–68.
- Piaget, J. (1952). *The origins of intelligence in children*. International Universities Press.
- Piaget, J. (1954). *The construction of reality in the child*. Basic Books.
- Piaget, J., & Inhelder, B. (1971). *Mental imagery in the child; a study of the development of imaginal representation*. Basic Books.
- Pinker, S., Choate, P. A., & Finke, R. A. (1984). Mental extrapolation in patterns constructed from memory. *Memory & Cognition*, 12(3), 207–218.
- Porro, C. A., Francescato, M. P., Cettolo, V., Diamond, M. E., Baraldi, P., Zuiani, C., Bazzocchi, M., & Di Prampero, P. E. (1996). Primary motor and sensory cortex activation during motor performance and motor imagery: A functional magnetic resonance imaging study. *Journal of Neuroscience*, 16(23), 7688–7698.
- Port, R. F., & Van Gelder, T. (Eds.). (1995). *Mind as motion: Explorations in the dynamics of cognition*. MIT Press.
- Posner, M. I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology*, 32(1), 3–25.

- Posner, M. I., Cohen, Y., Choate, L. S., Hockey, R., & Maylor, E. (1984). Sustained concentration: Passive filtering or active orienting? In *Preparatory States & Processes*. Psychology Press.
- Posner, M. I., Nissen, M. J., & Ogden, W. C. (1978). Attended and unattended processing modes: The role of set for spatial location. *Modes of Perceiving and Processing Information*, 137(158), 2.
- Powers III, A. R., Kelley, M., & Corlett, P. R. (2016). Hallucinations as top-down effects on perception. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 1(5), 393–400.
- Prinz, J. J. (2004). *Furnishing the mind: Concepts and their perceptual basis*. MIT press.
- Prinz, J. J., & Barsalou, L. W. (2000). Steering a course for embodied representation. In E. Dietrich & A. Markman (Eds.), *Cognitive dynamics: Conceptual and representational change in humans and machines* (pp. 51–77). Psychology Press.
- Pulvermüller, F. (2018). Neural reuse of action perception circuits for language, concepts and communication. *Progress in Neurobiology*, 160, 1–44.
- Qin, Y., & Simon, H. A. (1992). Imagery and mental models in problem solving. In N. H. Narayana (Ed.), *AAAI Spring Symposium on Reasoning with Diagrammatic Representations* (pp. 18–23). Stanford University.
- Qin, Z., Yu, F., Liu, C., & Chen, X. (2018). *How convolutional neural network see the world—A survey of convolutional neural network visualization methods*. arXiv. <https://doi.org/10.48550/arXiv.1804.11191>
- Ralph, M. A. L., Jefferies, E., Patterson, K., & Rogers, T. T. (2017). The neural and computational bases of semantic cognition. *Nature Reviews Neuroscience*, 18(1), 42–55.



- Ralph, M. A. L., Sage, K., Jones, R. W., & Mayberry, E. J. (2010). Coherent concepts are computed in the anterior temporal lobes. *Proceedings of the National Academy of Sciences, 107*(6), 2717–2722.
- Ramamurthy, U., Baars, B. J., D’Mello, S. K., & Franklin, S. (2006). LIDA: A working model of cognition. *Proceedings of the 7th International Conference on Cognitive Modeling*, 244–249.
- Ramamurthy, U., Negatu, A., & Franklin, S. (2001). Learning mechanisms for intelligent systems. *International Conference on Advances in Infrastructure for E-Business, e-Education and e-Science on the Internet. SSGRR-2001, L’Aquila, Italy.*
- Ratcliff, R. (1990). Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions. *Psychological Review, 97*, 285–308.  
<https://doi.org/10.1037/0033-295X.97.2.285>
- Reilly, J., Peelle, J. E., Garcia, A., & Crutch, S. J. (2016). Linking somatic and symbolic representation in semantic memory: The dynamic multilevel reactivation framework. *Psychonomic Bulletin & Review, 23*(4), 1002–1014.
- Rescorla, M. (2019). The language of thought hypothesis. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2019). Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/sum2019/entries/language-thought/>
- Rescorla, M. (2020). The computational theory of mind. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2020). Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/fall2020/entries/computational-mind/>
- Rieger, M., Dahm, S. F., & Koch, I. (2017). Inhibition in motor imagery: A novel action mode switching paradigm. *Psychonomic Bulletin & Review, 24*(2), 459–466.

- Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience*, 27(1), 169–192. <https://doi.org/10.1146/annurev.neuro.27.070203.144230>
- Rizzolatti, G., Fadiga, L., Gallese, V., & Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. *Cognitive Brain Research*, 3(2), 131–141.
- Rizzolatti, G., Riggio, L., Dascola, I., & Umiltá, C. (1987). Reorienting attention across the horizontal and vertical meridians: Evidence in favor of a premotor theory of attention. *Neuropsychologia*, 25(1, Part 1), 31–40. [https://doi.org/10.1016/0028-3932\(87\)90041-8](https://doi.org/10.1016/0028-3932(87)90041-8)
- Rizzolatti, G., & Sinigaglia, C. (2016). The mirror mechanism: A basic principle of brain function. *Nature Reviews Neuroscience*, 17, 757–765.  
<https://doi.org/10.1038/nrn.2016.135>
- Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In N. Navab, J. Hornegger, W. M. Wells, & A. F. Frangi (Eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (pp. 234–241). Springer International Publishing. [https://doi.org/10.1007/978-3-319-24574-4\\_28](https://doi.org/10.1007/978-3-319-24574-4_28)
- Rosch, E., & Lloyd, B. B. (Eds.). (1978). *Cognition and categorization*. Lawrence Erlbaum.
- Rosenbloom, P. S., Demski, A., & Ustun, V. (2016). The Sigma cognitive architecture and system: Towards functionally elegant grand unification. *Journal of Artificial General Intelligence*, 7(1), 1–103.
- Roth, W.-M., & Jornet, A. (2013). Situated cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, 4(5), 463–478.
- Russell, S., & Norvig, P. (2010). *Artificial intelligence: A modern approach* (3rd ed.). Prentice-Hall. (Original work published 1995)

- Sarker, M. K., Zhou, L., Eberhart, A., & Hitzler, P. (2021). *Neuro-symbolic artificial intelligence: Current trends*. arXiv. <https://doi.org/10.48550/arXiv.2105.05330>
- Scheil, J., & Liefoghe, B. (2018). Motor command inhibition and the representation of response mode during motor imagery. *Acta Psychologica, 186*, 54–62.
- Schiller, D., Eichenbaum, H., Buffalo, E. A., Davachi, L., Foster, D. J., Leutgeb, S., & Ranganath, C. (2015). Memory and space: Towards an understanding of the cognitive map. *Journal of Neuroscience, 35*(41), 13904–13911.
- Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). *Proximal policy optimization algorithms*. arXiv. <https://doi.org/10.48550/arXiv.1707.06347>
- Scott, S. H., & Kalaska, J. F. (2021). Voluntary movement: Motor cortices. In E. R. Kandel, J. D. Koester, S. H. Mack, & S. A. Siegelbaum (Eds.), *Principles of neural science* (6th ed., pp. 815–859). McGraw-Hill.
- Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences, 3*, 417–424.
- Sesmero, M. P., Ledezma, A. I., & Sanchis, A. (2015). Generating ensembles of heterogeneous classifiers using stacked generalization. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 5*(1), 21–34.
- Shanahan, M. (2006). A cognitive architecture that combines internal simulation with a global workspace. *Consciousness and Cognition, 15*, 433–449.
- Shanton, K., & Goldman, A. (2010). Simulation theory. *Wiley Interdisciplinary Reviews: Cognitive Science, 1*(4), 527–538. <https://doi.org/10.1002/wcs.33>
- Sharma, N., & Baron, J.-C. (2013). Does motor imagery share neural networks with executed movement: A multivariate fMRI analysis. *Frontiers in Human Neuroscience, 7*, Article 564. <https://doi.org/10.3389/fnhum.2013.00564>

- Shaver, P., Pierson, L., & Lang, S. (1974). Converging evidence for the functional significance of imagery in problem solving. *Cognition*, 3(4), 359–375.
- Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, 171(3972), 701–703.
- Singh, S. P., & Sutton, R. S. (1996). Reinforcement learning with replacing eligibility traces. *Machine Learning*, 22(1), 123–158.
- Slotnick, S. D., Thompson, W. L., & Kosslyn, S. M. (2012). Visual memory and visual mental imagery recruit common control and sensory regions of the brain. *Cognitive Neuroscience*, 3(1), 14–20.
- Smits-Engelsman, B. C., & Wilson, P. H. (2013). Age-related changes in motor imagery from early childhood to adulthood: Probing the internal representation of speed-accuracy trade-offs. *Human Movement Science*, 32(5), 1151–1162.
- Snaider, J., & Franklin, S. (2014a). Modular composite representation. *Cognitive Computation*, 6, 510–527.
- Snaider, J., & Franklin, S. (2014b). Vector LIDA. *Procedia Computer Science*, 41, 188–203.
- Snaider, J., McCall, R., & Franklin, S. (2011). The LIDA framework as a general tool for AGI. In J. Schmidhuber, K. R. Thórisson, & M. Looks (Eds.), *Artificial General Intelligence* (pp. 133–142). Springer Berlin Heidelberg.
- Souto, D. O., Cruz, T. K. F., Fontes, P. L. B., Batista, R. C., & Haase, V. G. (2020). Motor imagery development in children: Changes in speed and accuracy with increasing age. *Frontiers in Pediatrics*, 8, Article 100.
- Sowa, J. F. (Ed.). (2014). *Principles of semantic networks: Explorations in the representation of knowledge*. Morgan Kaufmann. (Original work published 1991)

- Spence, C., & Driver, J. (1994). Covert spatial orienting in audition: Exogenous and endogenous mechanisms. *Journal of Experimental Psychology: Human Perception and Performance*, 20(3), 555–574. <https://doi.org/10.1037/0096-1523.20.3.555>
- Spence, C., & Gallace, A. (2007). Recent developments in the study of tactile attention. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, 61(3), 196–207. <https://doi.org/10.1037/cjep2007021>
- Sperry, R. W. (1950). Neural basis of the spontaneous optokinetic response produced by visual inversion. *Journal of Comparative and Physiological Psychology*, 43(6), 482–489. <https://doi.org/10.1037/h0055479>
- Spruijt, S., van der Kamp, J., & Steenbergen, B. (2015). Current insights in the development of children's motor imagery ability. *Frontiers in Psychology*, 6, 787.
- Stanovich, K. E., & West, R. F. (2000). Advancing the rationality debate. *Behavioral and Brain Sciences*, 23(5), 701–717.
- Stramandinoli, F., Cangelosi, A., & Marocco, D. (2011). Towards the grounding of abstract words: A neural network model for cognitive robots. *The 2011 International Joint Conference on Neural Networks*, 467–474.
- Strang, G. (2016). *Introduction to linear algebra* (5th ed.). Wellesley-Cambridge Press.
- Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.
- Thagard, P. (2020). Cognitive science. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2020). Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/win2020/entries/cognitive-science/>
- Thelen, E., & Smith, L. B. (1994). *A dynamic systems approach to the development of cognition and action*. MIT press.

- Tian, X., Ding, N., Teng, X., Bai, F., & Poeppel, D. (2018). Imagined speech influences perceived loudness of sound. *Nature Human Behaviour*, 2(3), 225–234.
- Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review*, 55(4), 189–208.
- Toomela, A. (1999). A perceptual theory of knowledge: Specifying some details [Peer commentary on “Perceptual Symbol Systems,” by L. W. Barsalou]. *Behavioral and Brain Sciences*, 22(4), 633–634.
- Tyler, L. K., Stamatakis, E. A., Bright, P., Acres, K., Abdallah, S., Rodd, J. M., & Moss, H. E. (2004). Processing objects at different levels of specificity. *Journal of Cognitive Neuroscience*, 16(3), 351–362.
- Uhl, F., Goldenberg, G., Lang, W., Lindinger, G., Steiner, M., & Deecke, L. (1990). Cerebral correlates of imagining colours, faces and a map—II. negative cortical DC potentials. *Neuropsychologia*, 28(1), 81–93. [https://doi.org/10.1016/0028-3932\(90\)90088-6](https://doi.org/10.1016/0028-3932(90)90088-6)
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.
- Vogel, E. K., Woodman, G. F., & Luck, S. J. (2001). Storage of features, conjunctions, and objects in visual working memory. *Journal of Experimental Psychology: Human Perception and Performance*, 27(1), 92–114. <https://doi.org/10.1037/0096-1523.27.1.92>
- Von Holst, E. (1954). Relations between the central nervous system and the peripheral organs. *British Journal of Animal Behaviour*, 2, 89–94. [https://doi.org/10.1016/S0950-5601\(54\)80044-X](https://doi.org/10.1016/S0950-5601(54)80044-X)

- Waller, D., Schweitzer, J. R., Brunton, J. R., & Knudson, R. M. (2012). A century of imagery research: Reflections on Cheves Perky's contribution to our understanding of mental imagery. *The American Journal of Psychology*, *125*(3), 291–305.
- Wang, J., Conder, J. A., Blitzer, D. N., & Shinkareva, S. V. (2010). Neural representation of abstract and concrete concepts: A meta-analysis of neuroimaging studies. *Human Brain Mapping*, *31*(10), 1459–1468.
- Watkins, C. J., & Dayan, P. (1992). Q-learning. *Machine Learning*, *8*(3), 279–292.
- Watson, J. B. (1913). Psychology as the behaviorist views it. *Psychological Review*, *20*(2), 158–177. <https://doi.org/10.1037/h0074428>
- Wexler, M., Kosslyn, S. M., & Berthoz, A. (1998). Motor processes in mental rotation. *Cognition*, *68*(1), 77–94.
- Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin & Review*, *9*(4), 625–636.
- Wilson, S. W. (1991). The animat path to AI. In J.-A. Meyer & S. W. Wilson (Eds.), *From Animals to Animats: Proceedings of the First International Conference on Simulation of Adaptive Behavior* (pp. 15–21). MIT Press.
- Wintermute, S. (2012). Imagery in cognitive architecture: Representation and control at multiple levels of abstraction. *Cognitive Systems Research*, *19*, 1–29.
- Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, *5*(2), 241–259.
- Wolpert, D. M., Ghahramani, Z., & Jordan, M. I. (1995). An internal model for sensorimotor integration. *Science*, *269*(5232), 1880–1882.

- Xiao, H., Rasul, K., & Vollgraf, R. (2017). *Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms*. arXiv.  
<https://doi.org/10.48550/arXiv.1708.07747>
- Zacks, J. M. (2008). Neuroimaging studies of mental rotation: A meta-analysis and review. *Journal of Cognitive Neuroscience*, *20*(1), 1–19.
- Ziemke, T. (1999). Rethinking grounding. In *Understanding representation in the cognitive sciences* (pp. 177–190). Springer.
- Zwaan, R. A. (2004). The immersed experiencer: Toward an embodied theory of language comprehension. *Psychology of Learning and Motivation*, *44*, 35–62.
- Zwaan, R. A. (2014). Embodiment and language comprehension: Reframing the discussion. *Trends in Cognitive Sciences*, *18*(5), 229–234.
- Zwaan, R. A., Stanfield, R. A., & Madden, C. J. (1999). Perceptual symbols in language comprehension: Can an empirical case be made? *Behavioral and Brain Sciences*, *22*(4), 636–637.
- Zwaan, R. A., & Taylor, L. J. (2006). Seeing, acting, understanding: Motor resonance in language comprehension. *Journal of Experimental Psychology: General*, *135*(1), 1–11.  
<https://doi.org/10.1037/0096-3445.135.1.1>



## Appendix

### Representational Properties

This section reviews and attempts to analyze representational properties that are referenced throughout this manuscript. Many of these were ascribed by Barsalou (and others) to describe the flavor of non-symbolic representations used in perceptual symbol systems (Barsalou, 1999) and other conceptualizations of grounded cognition.

#### *Modal*

Barsalou (1999) stated, “the divergence between cognition and perception reflects the widespread assumption that cognitive representations are inherently nonperceptual, or what I will call *amodal*” (Barsalou, 1999, p. 577). In contrast, Barsalou defined *modal* representations as being “represented in the same systems as the perceptual states that produced them” (Barsalou, 1999, p. 578), and “perceptual states” as being composed from the *patterns of activation* occurring in sensory and motor systems during perception and action.

Barsalou argued that partial and attenuated versions of these perceptual states (comprising only their most salient features) could be learned into long-term memory from conscious experiences. Once learned, they could be recalled from long-term memory and function as grounded referents to entities, objects, and situations, allowing one to think about them in their absence (i.e., during offline cognitive activities). He referred to these re-enacted perceptual states as modal (or perceptual) symbols<sup>1</sup> (see Figure 37, Left Panel). This modal or

---

<sup>1</sup> Barsalou’s characterization of these modal representations as perceptual “symbols” is counter-intuitive, as they are not symbolic in the Peircian sense (see Peirce, 1867–1893/1992, pp. 225–228). I will try to avoid this terminology whenever possible, using the term modal (or perceptual) representations instead.

perceptual account of cognition contrasts with traditional amodal or non-perceptual accounts, which typically suggest that perceptual states are first transduced (i.e., mapped) into symbolic descriptions of those states, and it is those symbolic representations that support offline cognitive processes (see Figure 37, Right Panel).

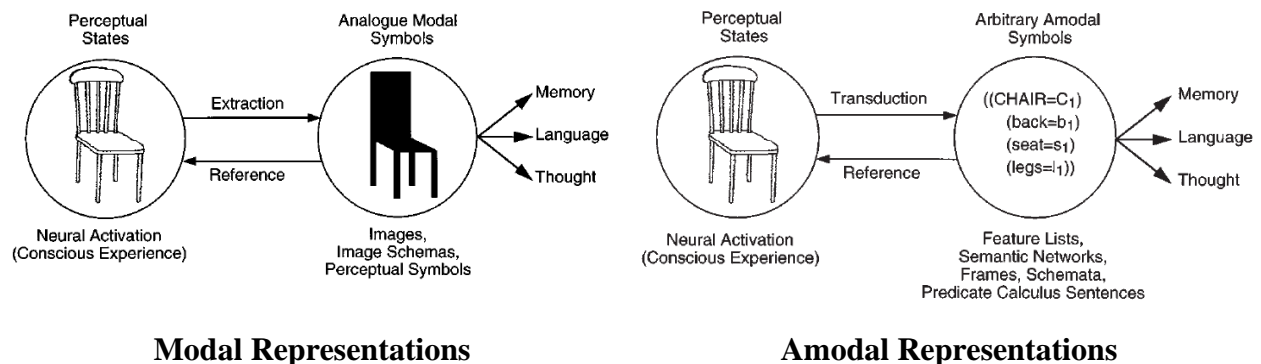


Figure 37. Conceptual depictions of modal and amodal representations. In both accounts, *perceptual states* arise from the activity in sensory and motor systems during perception and action. According to the modal account (Left Panel), the most salient aspects of these perceptual states are *extracted* to form modal representations. These can later serve as grounded referents to objects, entities, or situations during offline cognition. According to the traditional amodal account (Right Panel), perceptual states are *transduced* into amodal (non-perceptual) representations. These are manipulated using rule-based symbolic operations during offline cognition to generate new knowledge, and are mapped back to perceptual states (using conventional associations) to establish their referents. (Figures reproduced, without modification, from Barsalou, 1999.)

The adjective “modal” that characterizes modal representations stems from the idea that these representations originate in *modality-specific* sensory and motor systems. Prinz (2004) argued that the modality-specific nature of modal representations suggests that they may employ many disparate representational schemes that are proprietary to each modality. For example, he argued that visual and auditory sensory representations are likely *different in kind*. He argued this on the grounds that (1) sensory modalities have distinct inputs that may require different kinds of information processing, (2) sensory modalities are relatively independent of one another, (3) the phenomenal character of cognitive content originating in different sensory modalities is different (e.g., auditory, visual, and tactile sensations correspond to qualitatively different subjective

experiences), and (4) it is more “predictive and explanatory” of some experimental findings (e.g., timing effects in mental imagery tasks) to assume that sensory modalities employ different representational formats. (See Prinz, 2004, Chapter 5 for a more detailed treatment of these arguments).

This modal-specificity has led some to conclude that modal representations must respond to (be activated by) one, and only one, class of inputs. According to this view, representations capable of responding to inputs from multiple sensory or motor modalities are *amodal* by definition (e.g., see Dove, 2009; Machery, 2016; Reilly et al., 2016). Barsalou (2016a) disagreed. Instead, he contended that “multimodal abstractions” could exist that are responsive to inputs from multiple modalities while retaining the character of exemplars from individual modalities.

He gives two possible implementations. The first is based on *cross-modal conjunctive representations* (CCRs; see Binder, 2016) that store the “statistically likely features” extracted from category exemplars<sup>2</sup>. The second is multimodal compressed representations<sup>3</sup> resulting from the transformation of “exemplar information to a new set of dimensions (as in PCA, ICA, NMF, and so forth)” (Barsalou, 2016a, p. 1133). Barsalou included neural network models that use dimensional transformations to represent knowledge (e.g., autoencoders; see Chapter 5) in this latter group. Barsalou (2016a) characterized these multimodal compressions as *modal* because

---

<sup>2</sup> While Barsalou (2016a) does not mention it explicitly, these compressed multimodal abstractions would also be needed for *single* multimodal experiences (i.e., one exemplar) to bind together activity from multiple modalities. I will argue later that the use of mediating amodal representations as “associative hubs” leads to a simpler implementation that is still consistent with this “modality-specific” property.

<sup>3</sup> Barsalou (2016a) suggested that this could be accomplished using “multimodal compressions” that might exist in convergence zones (Damasio, 1989) in brains.

they contain information about modality-specific exemplars, allowing them to retain their modality-specific characteristics.

In summary, while modal-specificity implies that modal representations are responsive to inputs from a given modality, it does not rule out their responsiveness to multiple modalities. However, since amodal representations may also be responsive to multiple modalities, *input responsiveness cannot be used to differentiate modal from amodal representations*. More useful, in my opinion, is the view that modal-specificity simply implies that the activity in sensory and motor areas is constitutive of modal representations. As Barsalou (1999) stated, modal representations are formed from a subset of the sensory and/or motor activity corresponding to conscious experiences.

Therefore, I contend that modal representations are modality-specific because they are *necessarily* composed from the activity originating in sensory and/or motor systems. This criterion permits the inclusion of activity from one *or more* sensory and motor modalities. Representations that are not composed from sensory and/or motor patterns of activation are therefore amodal. Note that I am not claiming the *sufficiency* of this condition for declaring a representation modal. It is theoretically possible for an amodal representation to be partially composed of modality-specific content while violating other necessary conditions of modal representations (as outlined by Barsalou and others).

## *Analogical*

Barsalou (1999) stated that being modal implies that a representation is also *analogical*. He defined this as a partial structural correspondence<sup>4</sup> between a modal representation and the perceptual states that produced it (see Barsalou, 1999, p. 578). I interpret this to loosely imply that a “resemblance” exists between modal representations and their originating sensory and motor states that allows them to retain *some of the properties* implicit in those perceptual states. In other words, modal representations are *necessarily* non-symbolic, and more specifically, they are iconic representations in Peirce’s terminology (see Chapter 2).

Kugele and Franklin (2020b) further suggested that analogical representations must be capable of serving as “proxies” that reliably preserve the degree of similarity between any two perceptual states. More specifically, two analogical representations should be judged similar (by an agent’s perceptual processes) if and only if the saliency-modulated activation patterns of their corresponding perceptual states are similar. Critically, this property enables a modal cognitive system to *discriminate* (see Harnad, 1990, sec. 3.1) between patterns of sensory and motor activity based on their corresponding modal representations. Notice that analogical representations thus conceived bear a strong resemblance to Harnad’s *iconic representations*, which he defined as “analogs of the proximal sensory projections of distal objects and events” (Harnad, 1990, p. 335). Furthermore, contextualized abstractions over these analogical representations could generate more general and invariant analogical representations that

---

<sup>4</sup> One subtlety that bears mentioning is that Barsalou (1999) is not claiming with this property that a correspondence exists between modal representations and *their worldly referents*. He is only asserting that a structural correspondence must exist between modal representations and *their originating perceptual states* (i.e., sensory and motor activity). Specifically, he mentioned that the structure of perceptual symbols may correspond to the physical world in some cases, but not in others (see Barsalou, 1999, n. 1).

correspond roughly to Harnad's *categorical representations* (cf. also *cross-modal conjunctive representations* discussed earlier).

One remaining question is: "Are there amodal representations that satisfy the analogical property?" Machery (2006) claimed they can, stating

[there is a] large body of behavioral and neuropsychological evidence that humans have in fact *amodal, analogic representations of the approximate cardinality of classes of entities* (objects, sounds, events)... Adults as well as, to some extent, babies are able to estimate the approximate cardinality of classes of objects and to compare classes according to their cardinality, suggesting that they represent the approximate cardinality of classes.... Although there are several models of these representations, these models concur in regarding the representations of cardinality as *amodal* and *analogic*. (Machery, 2006, p. 406)

Furthermore, Prinz (2004) argued that if there were a "Language of Thought" (see Fodor, 1975, 2008) it could (at least theoretically) exhibit an isomorphic relationship to the things it represents (see Prinz, 2004, p. 112). Perhaps these amodal representations could be described as having the analogical property? Rather than further engage in this controversy here, I will simply assume in this manuscript that analogical, amodal representations could exist.

### ***Grounded***

Barsalou stated that while "mechanisms outside sensory-motor systems enter into conceptual knowledge, perceptual symbols *always* [emphasis added] remain grounded in these systems" (Barsalou, 1999, p. 583). Therefore, according to Barsalou, modal representations are *necessarily*

grounded. But what does “grounding” mean in the context of a purely non-symbolic and perceptual account of cognition? And what is its relationship with the intentionality (i.e., the “aboutness”) of a representation, and its intrinsic meaning?

A brief tour of the literature shows a multitude of, often contradictory, opinions and assumptions about grounding. For example, Harnad argued, with respect to his hybrid (symbolic/non-symbolic) architecture, that

[s]ymbolic representations must be *grounded bottom-up in nonsymbolic representations* [emphasis added] of two kinds: (1) *iconic representations*, which are analogs of the proximal sensory projections of distal objects and events, and (2) *categorical representations*, which are learned and innate feature detectors that pick out the invariant features of object and event categories from their sensory projections. (Harnad, 1990, p. 335)

For Harnad, establishing grounding is *equivalent* to establishing a symbolic representation’s “intrinsic meaning” or intentionality (see Harnad, 1990, sec. 2.1). Furthermore, Harnad argued that non-symbolic representations lack intrinsic meaning. He stated,

[i]conic representations no more “mean” the objects of which they are the projections than the image in a camera does. Both icons and camera images can of course be interpreted as meaning or standing for something, but the interpretation would clearly be derivative rather than intrinsic.... Nor can categorical representations yet be interpreted as “meaning” anything.... [they] do not have all the systematic properties of symbols.... They are just an inert taxonomy. (Harnad, 1990, p. 343)

Since Harnad (1990) equated grounding with intentionality (i.e., intrinsic meaning), and he considered non-symbolic representations meaningless, does that imply that he considered non-symbolic representations to be “ungrounded”? This seems unlikely, though a strict grounded or ungrounded representational dichotomy would compel him to accept this designation. Instead, Harnad might argue that the term simply does not apply in a non-symbolic context, or that non-symbolic representations are part of “the ground.” Support for this interpretation comes from Harnad’s statements regarding his iconic and categorical representations:

[t]here is no problem about [iconic and categorical representations] connection to the objects they pick out: It is a purely causal connection, based on the relation between distal objects, proximal sensory projections and the acquired internal changes that result from a history of behavioral interactions with them. (Harnad, 1990, p. 343)

Therefore, Harnad is not suggesting a disconnect between these non-symbolic representations and their world referents, only that these causal connections are insufficient for intrinsic meaning.

Now that I have summarized Harnad’s positions on grounding and intentionality, let us look at a non-representational account. Recall Brooks’s “physical grounding hypothesis” (see Chapter 2), which entailed building systems that are *grounded in the physical world*. Brooks stated,

[a]ccepting the physical grounding hypothesis as a basis for research entails building systems in a *bottom up manner* [emphasis added]. High level abstractions have to be made concrete. The constructed system eventually has to express all its goals and desires



as physical action, and *must extract all its knowledge from physical sensors* [emphasis added].... The forms of the low-level interfaces have consequences which ripple through the entire system. (Brooks, 1990, p. 5)

For Brooks, this principle manifested itself in the building of intelligent systems whose (re)actions were directly plugged into their sensory inputs without intervening representations. That said, he was not arguing that grounded systems are necessarily non-representational, only that “it is necessary to have [their] representations grounded *in the physical world* [emphasis added]” (Brooks, 1990, p. 5). Physical grounding, according to Brooks, requires connecting a system *to the world* via a set of sensors and actuators. He goes on to state that “typed input[s] and output[s] are no longer of interest,” as they are not physically grounded.

Finally, let me circle back and complete my summary of Barsalou’s notion of grounding. Recall that Barsalou’s modal representations (i.e., the things to be grounded) are a types of non-symbolic representations per Peirce’s semiotics, though Barsalou choose to call them (perceptual or modal) “symbols.” Furthermore, Barsalou stated that modal representations are *grounded in sensory and motor systems*, not non-symbolic representations (like Harnad’s symbol grounding), and not the physical world (like Brooks’s physical grounding). To illustrate this in the context of an implementation, Barsalou gave the example of a *shared* associative (neural) network: “If the same associative network represents information in both perception and cognition, it grounds knowledge in perception” (Barsalou, 1999, p. 579). In other words, Barsalou’s notion of grounding is focused on eliminating potential disconnects between perception and cognition through representational (e.g., neural) reuse. That said, Barsalou does not deny the existence of

other grounding mechanisms and grounding contexts, such as physical, social, and bodily grounding (Barsalou, 2020, sec. 1.1).

Unlike Harnad, Barsalou distinguishes between grounding and intentionality. For example, he stated that (1) the content of a representation does not specify its intentionality, (2) the degree to which a representation resembles its referent is neither sufficient nor necessary for establishing reference, and (3) factors external to a representation's content play an important role in establishing its intentionality (see Barsalou, 1999, p. 597). Given that modal representations are always grounded, initially characterized only by their content, and that content alone does not fully specify a representation's intentionality, it follows that, according to Barsalou, modal representations can be grounded, yet lack intrinsic meaning. That is, intentionality must be distinct from grounding in Barsalou's theory. That said, Barsalou argued that modal representations have an advantage over symbolic representations in establishing their intentionality, as their content can *heuristically* assist in establishing reference (see Barsalou, 1999, sec. 3.2.8).

It is remarkable that in the three accounts of grounding described above nearly everything about the specifics of these accounts differ. Let us review.

*What is being ground?*

- symbolic representations (Harnad)
- non-symbolic representations (Barsalou)
- intelligent systems (Brooks)

*What are they grounded in?*

- two kinds of non-symbolic representations (Harnad)
- sensory and motor systems (Barsalou)
- the physical world (Brooks)

Furthermore, they all embrace different notions of the relationship between grounding and intentionality. Harnad (1990) suggested an equivalence between grounding and intentionality. Barsalou (1999) argued that they were distinct. And Brooks seems to have dismissed the problem. The one commonality is that all of these researchers explicitly state or imply that grounding requires *bottom-up* construction of representations and systems, which I disagree with—e.g., “eventual grounding” (see Chapter 3) entails the creation of an amodal symbol for a concept or concept instance prior to its grounding.

How do we reconcile these differences? My contention is that the only way to unify these accounts is to start by separating intentionality from grounding as Barsalou does. According to the Stanford Encyclopedia of Philosophy, intentionality is defined as

the power of minds and mental states to be about, to represent, or to stand for, things, properties and states of affairs. To say of an individual’s mental states that they have intentionality is to say that they are mental representations or that they have contents.

(Jacob, 2020, para. 1)

Consequentially, conflating grounding and intentionality immediately threatens to alienate non-representational accounts of cognition (e.g., Brooks’s subsumption architecture). Second, if establishing intrinsic meaning (i.e., intentionality) is synonymous with establishing grounding, then any representation or thing that is not *the thing being grounded* (i.e., being imbued with

intrinsic meaning) seems to deserve the characterization of “ungrounded.” For example, Harnad’s declaration that non-symbolic representations are “meaningless” seems to force one to label them as ungrounded in his theory. Barsalou would similarly have to label modal representations that lack intentionality as “ungrounded,” which is a contradiction to his assertion that they are *always* grounded.

Rather than appealing to vague or intuitive notions of grounding, it will be beneficial for the purpose of this thesis to provide a working definition of what I mean by grounding. This definition is intended to serve as a means of qualifying the presence of representational or systemic grounding in an engineered system, and to be inclusive of the pertinent aspects of all three notions of grounding previously described. Ironically, Harnad’s characterization of non-symbolic representations (which he regarded as “meaningless”) served as the basis for my working definition. Namely, I propose the following working definition:

**grounding** is the establishment of information-bearing associations between sensory and motor stimuli, and the representations and/or processes that use them.

These informational conduits allow the states of the former to induce, and covary with, the states (e.g., patterns of activation) of the latter. For example, the proximal sensory projections (e.g., retinal projections) resulting from worldly objects can induce, and covary with, internal mental states<sup>5</sup>.

---

<sup>5</sup> This working definition is inclusive of internally derived sensory and motor stimuli (e.g., mental simulations), as well as both innate and acquired informational connections.

While distinct from intentionality, these informational conduits support an agent's often-gradual, life-long acquisition of intrinsic meaning. In other words, grounding is the means by which intentionality can be experientially discovered. Ziemke's (1999) interpretation of the *grounding problem* provides a nice example of how my conception of grounding and the one typically assumed in the literature relate. He stated,

[t]he grounding problem is, generally speaking, the problem of how to causally connect an artificial agent with its environment such that the agent's behaviour, as well as the mechanisms, representations, etc. underlying it, can be intrinsic and meaningful to itself, rather than dependent on an external designer or observer. (Ziemke, 1999, p. 177)

The version of the grounding problem I adopt here only corresponds to "how to causally connect an artificial agent with its environment." I would characterize the remainder of Ziemke's definition as concerning the problem of intentionality or intrinsic meaning.

In summary, modal representations are *necessarily* grounded, where grounding involves the establishment of causal connections between an agent and its (external or internal) environment. The existence of these causal connections does not imply the intentionality of internal states. Intentionality and grounding are distinct. However, grounding is instrumental of the acquisition of intrinsic meaning through experience. With respect to Brooks's subsumption architecture, *physical grounding* involves the hardwiring of (innate/built-in) causal connections between an agent's proximal sensory projections (via its sensors) and the internal states of its behavior modules. Note that this assumes that the proximal sensory projections occurring in the agent's sensors are grounded in the world; otherwise, the agent's sensors would be useless, and the system would not be physically grounded. Barsalou's *sensorimotor grounding* involves the

establishment of causal connections between an agent's sensory and motor systems, and its (typically learned) modal representations. These connections allow modal representations to covary with the patterns of activation in sensory and motor systems. Note that this assumes some degree of grounding between the agent's sensory and motor systems, and its environment (potentially mediated by proximal sensory projections). Finally, with respect to Harnad's hybrid architecture, *symbol grounding* involves the establishment of causal connections between two kinds of non-symbolic representations and the symbolic representations to which they are connected. These connections take the form of associations that allow symbolic representations to acquire the covariance properties of their associated non-symbolic representations (e.g., being activated when their associated non-symbolic representations are activated). However, this assumes that the system's non-symbolic representations are already grounded in proximal sensory projections, which are, in turn, grounded in the world.

In the above descriptions, I have taken for granted the *transitivity* of grounding operations (i.e., if *X is grounded in Y*, and *Y is grounded in Z*, then *X is grounded in Z*). For example, this transitivity of grounding is what allows Harnad's symbolic representations to be *grounded in the world* via associations with non-symbolic representations that are grounded in sensory projections that are grounded in the world.

### ***Shared***

Perhaps the most distinctive characteristic of modal theories of cognition is the *reuse* of modality-specific sensory and motor resources for conceptual processing. Haimovici stated, "whichever properties modal representations exhibit, what crucially distinguishes modal from amodal approaches is a commitment to the constitutive involvement of sensorimotor systems in

conceptual tasks” (Haimovici, 2018, p. 7). Barsalou (2016a) supported this characterization, stating that the representational format used by sensory and motor systems is largely irrelevant. What matters is that conceptual processes depend on modality-specific representations and the perceptual processes that operate on them. Furthermore, the conceptual representations resulting from these processes always have “a modality-specific character, not an amodal one” (Barsalou, 2016a, p. 1131). Recall that Barsalou considered a *shared* neural network that supports both perceptual and conceptual processes to be a natural implementation of these ideas (Barsalou, 1999, p. 579).

By contrast, amodal theories of cognition often limit the involvement of perceptual systems to the transduction of sensory stimuli into symbolic representations. Moreover, conceptual knowledge is conceived of as being largely independent of sensory and motor systems. Consequently, perceptual processes and representations are not considered to be directly constitutive of conceptual processing. Furthermore, any conceptual representations resulting from amodal “thought” processes are considered amodal in character.

In summary, modal representations are *necessarily* shareable between, usable by, and constitutive of both perceptual and conceptual systems in modal approaches. As Barsalou stated over 20 years ago, a unifying idea of modal approaches to cognition is that “cognition is inherently perceptual” (Barsalou, 1999, p. 577). Therefore, representations incompatible with, or segregated from, a system’s perceptual system must be characterized as amodal.

### ***Embodied***

The idea that modal representations are “embodied” and amodal representations are “disembodied” has been widely embraced in the simulation-based cognition literature (e.g., see

Barsalou, 1999, sec. 3.3; De Vega et al., 2012; Dove, 2011, 2016; Gallese, 2018; Goldman, 2012, 2013; Mahon & Caramazza, 2008; Mahon & Hickok, 2016; Prinz & Barsalou, 2000; Pulvermüller, 2018). Unfortunately, the broader embodied cognition (EC) community has been less receptive. The first complication arises from a strong anti-representational sentiment that exists within the EC community. Consequently, claims of “embodied representations” will naturally seem foreign and unpalatable to many. The second complication results from an ongoing identity crisis within the EC literature: the basic terminology and core tenets are still in flux. The 4Es (embodied, embedded, enacted, and extended; see Newen et al., 2018), the “Six Views” (see M. Wilson, 2002), ecological cognition, situated cognition, and grounded cognition are all broadly aligned with EC, and yet they are distinct from it.

One might presume that declaring a species of mental representations “embodied” would entail providing an account of how bodies and bodily experiences shape the format, content, and systemic function of those representations. However, I think this accounting has been broadly achieved by the proponents of simulation-based theories of cognition. Simulation-based cognition offers a representational account of cognition based on the re-enactment of patterns of activation in sensory and motor systems. These patterns of activation directly reflect the idiosyncrasies of the body, and the content of those systems necessarily reflect bodily experiences. Any generalizations and abstractions over those perceptual states are constrained by these theories to maintain a modal character. And in its pure form, simulation-based cognitive theories eschew all things amodal. It is hard to imagine how a representational account of cognition could be more body-based.



Barsalou and other proponents of simulation-based theories of cognition have devoted a great deal of energy to differentiating simulation-based cognition from traditional cognitivist theories (e.g., the computational theory of mind; see Chapter 2). However, those attempts have not convinced everyone. For example, Gallagher (2015) characterized simulation-based theories (e.g., Barsalou, 1999, 2008; Goldman, 2013), as well as theories based on metaphorical thought (e.g., Lakoff & Johnson, 1980/2008), as “body snatchers.” Gallagher stated,

It’s clear that *the body* of this version of embodied cognition is entirely *in the head*; it’s the “body in the brain”. In this respect it is a “minimal” or “weak” form of embodied cognition, at best. A form of embodied cognition without the body as such... (Gallagher, 2015, p. 99)

Being “about the body” and having representations that are “based on the body” are inadequate justifications for “embodied” status according to Gallagher.

In defense of the embodiment of simulation-based theories of cognition, Barsalou proposed the idea of *variable embodiment*, which asserts that a representation’s meaning reflects the idiosyncrasies of the physical system that represents it. Several examples of variable embodiment were given by Zwaan et al. (1999) in the context of language comprehension. They stated,

it might be that a [large] basketball player represents the situation described by “He picked up the basketball” differently (one-handed) from a smaller person (two-handed). One would also assume that women who have given birth construct different mental

representations of a story about childbirth than other women or than men. (Zwaan et al., 1999, p. 636)

Barsalou later expanded on these ideas, contending that modal cognitive systems utilize the environment and the body as “external informational structures that complement internal representations.” He also stated that these internal representations take on a “situated character” when simulated in sensory and motor systems (see Barsalou, 2010, p. 717). As such, modal mental simulations have been variously referred to as *embodied simulations* (Gallese, 2005, 2007, 2018), *situated simulations* (Barsalou, 2009), and *situated conceptualizations* (Barsalou, 2009, 2016b).

In other words, Barsalou argued that modal cognitive systems are “embodied” in the sense that the idiosyncrasies of an agent’s body become constitutive of an agent’s conceptual representations, and, as a result, if two agents experience the world differently, these experiential differences manifest as representational differences in their respective conceptual systems (i.e., variable embodiment). However, modal representations are not embodied in the stricter sense used by Gallagher that an agent’s non-brain body parts must be directly engaged during cognitive activities.

Assuming one agrees that modal accounts of cognition feature embodied (and potentially situated) representations, does that preclude embodied amodal representations? Both Barsalou (1999) and Zwaan et al. (1999) argued that variable embodiment is impossible in amodal cognitive systems. Barsalou's argument appeals to the non-analogical nature of “word-like” symbolic representations. For example, he argued that the semantic interpretation of the word “CUP” is identical regardless of any variability in its representational content:

Making “CUP” larger does not mean that its referent now appears larger. Rotating “CUP” 45° counterclockwise does not imply that its referent tipped... Because words bear no structural relations to their referents, structural changes in a word have no implications for analogous changes in reference. As long as the conventional link between a word and its referent remains intact, the word refers to the referent discretely in exactly the same way across transformations. (Barsalou, 1999, p. 598)

Barsalou argued that modal representations are fundamentally different. For example, increasing the “size” of a modal representation corresponding to a cup might imply that its environmental referent appears larger in an individual’s visual field. This might happen, for example, when an agent moves closer to the cup. This transformed modal representation of that cup is not functionally equivalent to its untransformed counterpart.

However, Barsalou’s and Zwann’s arguments against amodal representations exhibiting variable embodiment are not entirely convincing. For example, consider an amodal cognitive system based on distributional semantics (see Chapter 2). Such cognitive systems interpret their internal “word-like” symbols based on the contexts in which their referents occurred. These contexts are based on an agent’s experiences, and those experiences are influenced by its body. Therefore, the meanings associated with those internal representations will inevitably differ between agents with different bodies and life experiences. Second, this argument neglects to consider composite symbolic representations that are capable of further qualifying sensory experiences—for example, “the tiny cup,” “the large cup,” and “the massive cup.” Even if one assumes that an agent’s conceptual system is amodal and fixed, it does not follow that the

semantic knowledge of such agents must, in principle, be agnostic of the agent's body.  
Therefore, I contend that amodal representations could exhibit variable embodiment.

## Learning Intelligent Decision Agent (LIDA)

Table 6. LIDA's short-term and long-term memory (STM/LTM) modules and codelets.

Module / Process	Description
ACTION SELECTION (AS)	STM module supporting the selection of behaviors (i.e., instantiated schemes) for execution by the SMS. Action Selection, in conjunction with other LIDA modules and processes, supports four modes of action selection: consciously mediated, volitional, automatized, and alarms.
ATTENTION CODELETS (ACs)	<p>Specialized processors that monitor the CSM for content of interest based on their own specific concerns, such as importance, urgency, novelty, etc. If such content is found, the codelet takes it to a coalition forming process, which may create a coalition that includes that codelet and the content it promotes.</p> <p>Attention codelets vary in the kinds of content they consider salient. <i>Specific attention codelets</i> advocate for a narrow range of content, for example, specific types of objects or events. By contrast, the <i>default attention codelet</i> advocates for a wide range of content: its selection criteria are based solely on the content's total activation and total incentive salience. <i>Expectation codelets</i> are attention codelets created in response to selected behaviors that advocate for content in the CSM corresponding to the expected results (or non-results) of that selected behavior.</p>
CONSCIOUS CONTENTS QUEUE (CCQ)	STM submodule of the Workspace that contains recent conscious broadcasts.
CURRENT SITUATIONAL MODEL (CSM)	STM submodule of the Workspace that represents an agent's (preconscious) interpretation of its current situation.
GLOBAL WORKSPACE (GW)	STM module that directs a winner-take-all competition among <i>coalitions</i> , and broadcasts the content of the winning coalition in the global (conscious) broadcast.

---

MOTOR PLAN EXECUTION (MPE)	See SMS.
PERCEPTUAL ASSOCIATIVE MEMORY (PAM)	LTM module that supports LIDA’s ability to recognize objects, events, entities, concepts, etc., and the relationships between them. The most activated representations in PAM are instantiated into the CSM as <i>percepts</i> after being activated by incoming sensory content (or cueing).
PROCEDURAL MEMORY (PM)	LTM module containing representations called <i>schemes</i> that each encode a context, action, and expected result. When schemes are instantiated (that is, when their free variables are bound to specific values based on the contents of a conscious broadcast) they are referred to as (candidate) <i>behaviors</i> .
SENSORY MEMORY (SM)	STM module that encodes modality-specific sensory content (from the environment) as the activation of low-level features detectors. These, in turn, activate perceptual representations in PAM. SM also sends sensory representations, based on the activation of its low-level feature detectors, to the CSM.
SENSORY MOTOR MEMORY (SMM)	See SMS.
SENSORY MOTOR SYSTEM (SMS)	Composed of two modules: Sensory Motor Memory and Motor Plan Execution. The SMS selects and instantiates motor plan templates from SMM into concrete motor plans, and sends them to the Motor Plan Execution module for execution.
STRUCTURE BUILDING CODELETS (SBCs)	Specialized processors that create or modify content in the CSM in support of “preconscious thought” and situational understanding.
WORKSPACE	STM module supporting preconscious, situational understanding. At any given moment it may contain cued long-term memories, percepts, sensory content (both real and simulated), transient representations created by structure building codelets. It contains two submodules—the CSM and CCQ.

---

## K-Armed Bandit Agent Implementation: Parameters and their Values

Table 7. Implementation parameters and their values ( $k$ -armed bandit agent).

Parameter	Value	Description
<i>learning rate</i>	0.001	Controls the rate of base-level incentive salience and reliability updates (among other things).
<i>composite actions chain maximum length</i>	3	Specifies the maximum length of any chain of schemes in a composite action's controller.
<i>composite actions minimum baseline advantage</i>	0.4	Specifies the minimum combined affective valence and incentive salience (over a learned baseline value) necessary to consider a node as a goal state for a new composite action. <sup>6</sup>
<i>composite action update frequency</i>	0.01	The probability (per broadcast) that a composite action's controller will be updated.
<i>epsilon decay rate</i>	0.9999	The (geometric) decay rate used for the random exploratory temperature.
<i>epsilon initial</i>	0.9999	The initial random exploratory temperature.
<i>epsilon minimum</i>	0.025	The minimum random exploratory temperature.
<i>base-level incentive salience eligibility trace discount factor</i>	0.5	The (geometric) value reduction applied for each time step between the occurrence of a node in a conscious broadcast and a later broadcast containing affective valence. (This was used to modulate base-level incentive salience updates.)
<i>base-level incentive salience eligibility trace decay rate</i>	0.5	The (geometric) decay rate applied to the eligibility trace value of each non-occurring node in a conscious broadcast. (This was used to modulate base-level incentive salience updates.)
<i>positive correlation threshold</i>	0.9	The positive correlation value used to determine when a spin-off occurred.
<i>pending focus decay rate</i>	0.4	The (geometric) decay rate used for the pending schemes focus bonus in Action Selection.
<i>pending focus max. value</i>	0.9	The maximum additional selection importance given to components of pending schemes.
<i>reliability max. penalty</i>	0.6	The maximum reduction in the selection importance of unreliable schemes.
<i>reliability threshold</i>	0.8	The minimum base-level activation for inclusion in a composite action.

---

<sup>6</sup> Drescher (1991, p. 90) suggested defining a composite action for every novel result; that is, every combination of known items would correspond to a goal state. However, for most environments this will be computationally intractable. In practice, goal states must be limited to avoid a proliferation of composite actions.

Table 8. Learned and built-in schemes for  $k$ -armed bandit agent ( $k = 8$ ). Schemes are listed in the order in which they were added to Procedural Memory. Base-level activations are provided for each non-bare scheme. Composite actions are shown in bold font.

<b>n</b>	<b>Schemes</b>	<b>Base-Level Act.</b>	<b>n</b>	<b>Schemes</b>	<b>Base-Level Act.</b>
1	/stand/	nan	41	/deposit/M6, P	0.26
2	/sit (M0) /	nan	42	M1, P/play/L	0.91
3	/sit (M3) /	nan	43	M7, P/play/W	1.00
4	/sit (M1) /	nan	44	/deposit/M7, P	0.46
5	/sit (M2) /	nan	45	/deposit/M0, P	0.08
6	/sit (M6) /	nan	46	M5, P/play/W	0.73
7	/sit (M4) /	nan	47	/deposit/M5, P	0.03
8	/play/	nan	48	M6/deposit/M6, P	1.00
9	/sit (M5) /	nan	49	/deposit/M1, P	0.03
10	/sit (M7) /	nan	50	M0/deposit/M0, P	1.00
11	/deposit/	nan	51	S/sit (M2) /M2	1.00
12	/sit (M0) /M0	0.20	52	M2, P/play/L	0.76
13	/stand/S	1.00	53	/deposit/M2, P	0.04
14	/sit (M3) /M3	0.25	54	M2, P/play/W	0.24
15	/sit (M5) /M5	0.33	55	M5/deposit/M5, P	1.00
16	/deposit/P	0.99	56	M7/deposit/M7, P	1.00
17	/sit (M1) /M1	0.25	57	M5, P/play/L	0.27
18	S/sit (M5) /M5	1.00	58	M2/sit (M2) /M2	1.00
19	/sit (M2) /M2	0.15	59	<b>/w/w</b>	0.75
20	M5/sit (M5) /M5	1.00	60	M2/deposit/M2, P	1.00
21	/sit (M7) /M7	0.62	61	M1/deposit/M1, P	1.00
22	S/sit (M1) /M1	1.00	62	M3, P/play/W	0.43
23	/sit (M6) /M6	0.55	63	/deposit/M3, P	0.03
24	/play/W	0.64	64	/sit (M4) /M4	0.27
25	<b>/w/</b>	nan	65	M3/deposit/M3, P	1.00
26	M1/sit (M1) /M1	1.00	66	M3, P/play/L	0.57
27	S/sit (M0) /M0	1.00	67	S/sit (M4) /M4	1.00
28	/play/L	0.17	68	M6, P/play/L	0.10
29	S/sit (M3) /M3	1.00	69	M4/sit (M4) /M4	1.00
30	P/play/L	0.21	70	M4, P/play/L	0.42
31	M3/play/W	0.20	71	/deposit/M4, P	0.06
32	M0/sit (M0) /M0	1.00	72	M4/deposit/M4, P	1.00
33	M3/sit (M3) /M3	1.00	73	<b>/w/L</b>	0.00
34	S/sit (M6) /M6	1.00	74	M3/ <b>w/W</b>	0.75
35	P/play/W	0.79	75	P/ <b>w/W</b>	0.93
36	S/sit (M7) /M7	1.00	76	M4, P/play/W	0.59
37	M0, P/play/L	1.00	77	M1, P/play/W	0.08
38	M7/sit (M7) /M7	1.00	78	M3, P/ <b>w/W</b>	0.92
39	M6/sit (M6) /M6	1.00	79	<b>/w/M7</b>	0.47
40	M6, P/play/W	0.91	80	M7/ <b>w/M7</b>	1.00



Table 9. Learned base-level incentives saliences for  $k$ -armed bandit agent ( $k = 8$ ).

<b>Base-Level Incentive Saliency</b>	<b>Node</b>
0.253	M7, P
0.187	M6, P
0.111	M7
0.107	P
0.077	M6
0.024	M5, P
0.009	M4, P
-0.001	S
-0.005	M5
-0.007	W
-0.030	M3, P
-0.044	M4
-0.056	M3
-0.101	M2, P
-0.107	M1, P
-0.114	M2
-0.124	M1
-0.174	L
-0.240	M0, P
-0.247	M0

## Simulation-Based Agent Implementation: $\beta$ -VAE Network Architecture

Layer (type)	Output Shape	Param #	Connected to
input_1 (InputLayer)	[(None, 128, 128, 1)]	0	
encoder (Encoder)	((None, 16), (None, 16))	912160	input_1[0][0]
sampler (Sampler)	(None, 16)	0	encoder[0][0] encoder[0][2]
decoder (Decoder)	(None, 128, 128, 1)	1256193	sampler[0][0]
recon_loss (ReconstructionLoss)	()	0	encoder[0][3] decoder[0][0]
kl_loss (KLLoss)	()	1	encoder[0][0] encoder[0][1]

=====  
 Total params: 2,168,354  
 Trainable params: 2,168,353  
 Non-trainable params: 1

Figure 38. Overview of  $\beta$ -VAE neural network architecture.

Layer (type)	Output Shape	Param #	Connected to
input_2 (InputLayer)	(None, 128, 128, 1)	0	
preprocessor (Preprocessor)	(None, 128, 128, 1)	0	input_2[0][0]
encoder/conv1 (Conv2D)	(None, 128, 128, 32)	320	preprocessor[0][0]
encoder/maxpool1 (MaxPooling2D)	(None, 64, 64, 32)	0	encoder/conv1[0][0]
encoder/conv2 (Conv2D)	(None, 64, 64, 64)	18496	encoder/maxpool1[0][0]
encoder/maxpool2 (MaxPooling2D)	(None, 32, 32, 64)	0	encoder/conv2[0][0]
encoder/conv3 (Conv2D)	(None, 32, 32, 128)	73856	encoder/maxpool2[0][0]
encoder/maxpool3 (MaxPooling2D)	(None, 16, 16, 128)	0	encoder/conv3[0][0]
encoder/conv4 (Conv2D)	(None, 16, 16, 256)	295168	encoder/maxpool3[0][0]
encoder/maxpool4 (MaxPooling2D)	(None, 8, 8, 256)	0	encoder/conv4[0][0]
encoder/flatten (Flatten)	(None, 16384)	0	encoder/maxpool4[0][0]
encoder/logvar (Dense)	(None, 16)	262160	encoder/flatten[0][0] encoder/flatten[0][0]
encoder/mu (Dense)	(None, 16)	262160	encoder/flatten[0][0]
encoder/sigma (Lambda)	(None, 16)	0	encoder/logvar[1][0]

Total params: 912,160  
 Trainable params: 912,160  
 Non-trainable params: 0

Figure 39. Overview of  $\beta$ -VAE encoder.

Layer (type)	Output Shape	Param #
input_18 (InputLayer)	[(None, None, 16)]	0
decoder/dense1 (Dense)	multiple	278528
decoder/reshape1 (Reshape)	(None, 8, 8, 256)	0
decoder/deconv1 (Conv2DTrans)	(None, 16, 16, 256)	590080
decoder/deconv2 (Conv2DTrans)	(None, 32, 32, 128)	295040
decoder/deconv3 (Conv2DTrans)	(None, 64, 64, 64)	73792
decoder/deconv4 (Conv2DTrans)	(None, 128, 128, 32)	18464
decoder/deconv5 (Conv2DTrans)	(None, 128, 128, 1)	289

Total params: 1,256,193  
 Trainable params: 1,256,193  
 Non-trainable params: 0

Figure 40. Overview of  $\beta$ -VAE decoder.

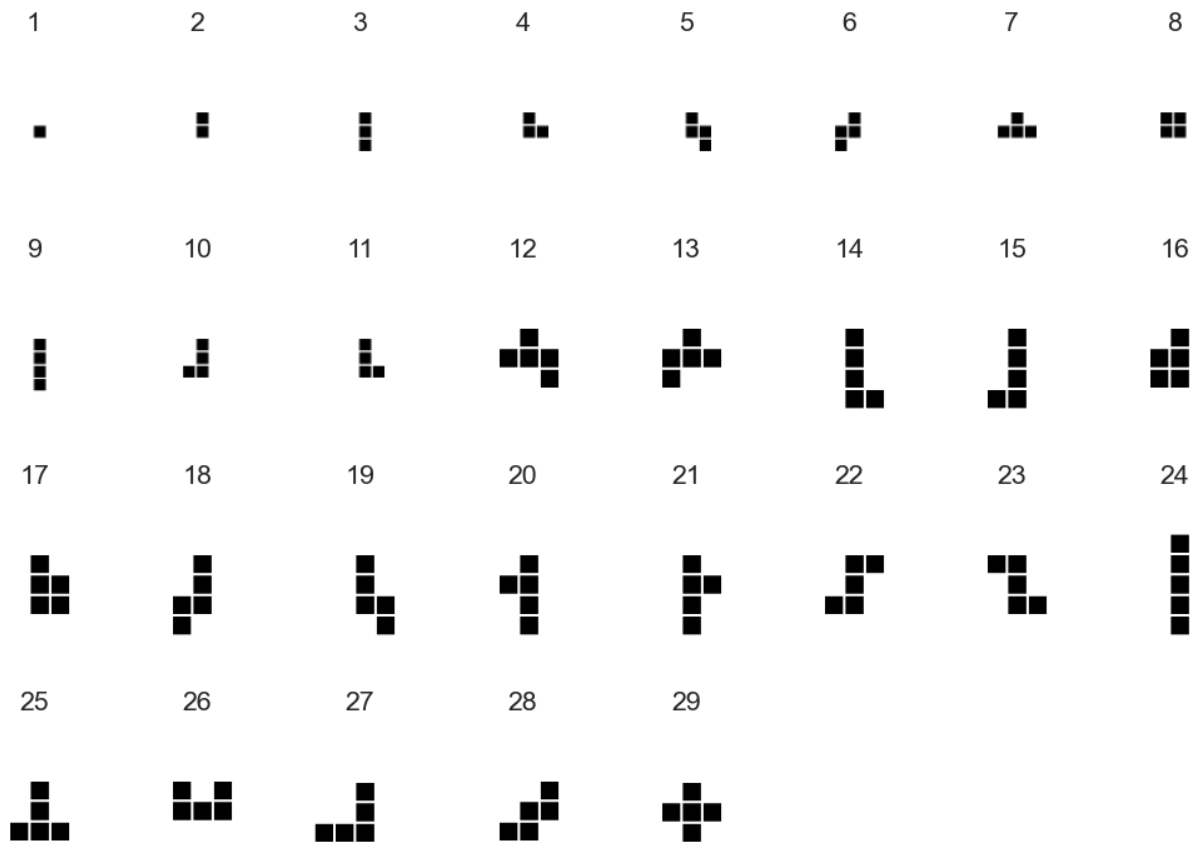


Figure 41. Polyominoes used in mental imagery environment. (Shown in standardized orientations).

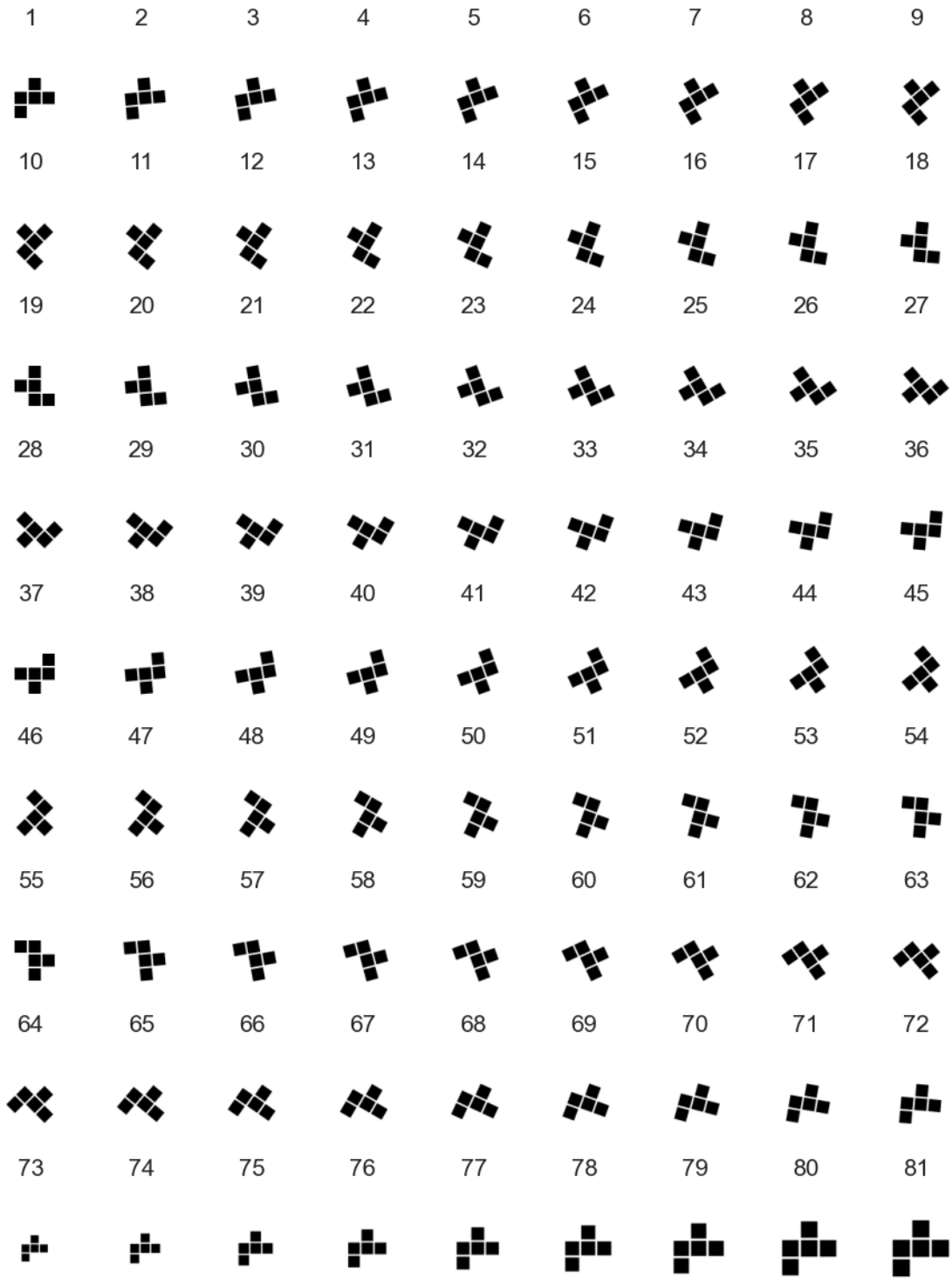


Figure 42. All rotations and scales for a single pentomino shape. (Shape 13).

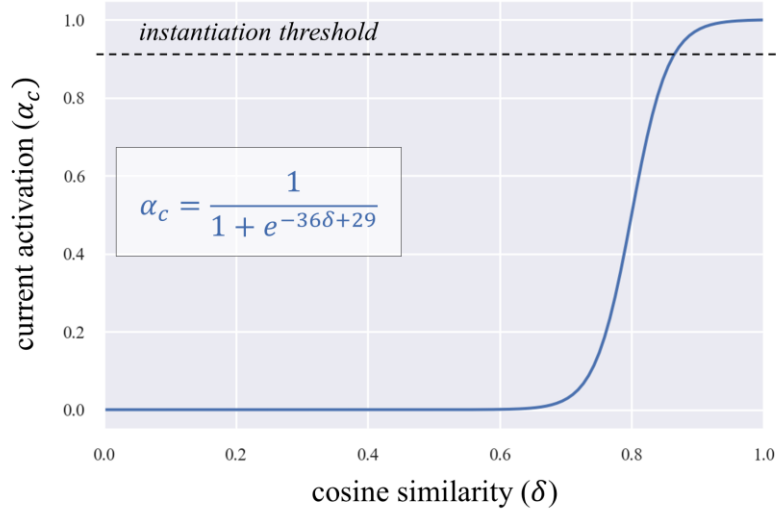


Figure 43. A sigmoidal current-activation function. The magnitude of a primitive feature detector’s *current activation* is based on the scaled (cosine) similarity between the modal probability distributions for incoming sensory stimuli and the previously learned modal probability distributions associated with primitive feature detector. PAM’s instantiation threshold is also depicted, which corresponds to the amount of *total activation* (current + base-level activation) that is needed for a percept to be instantiated into the LIDA agent’s Current Situational Model (CSM).

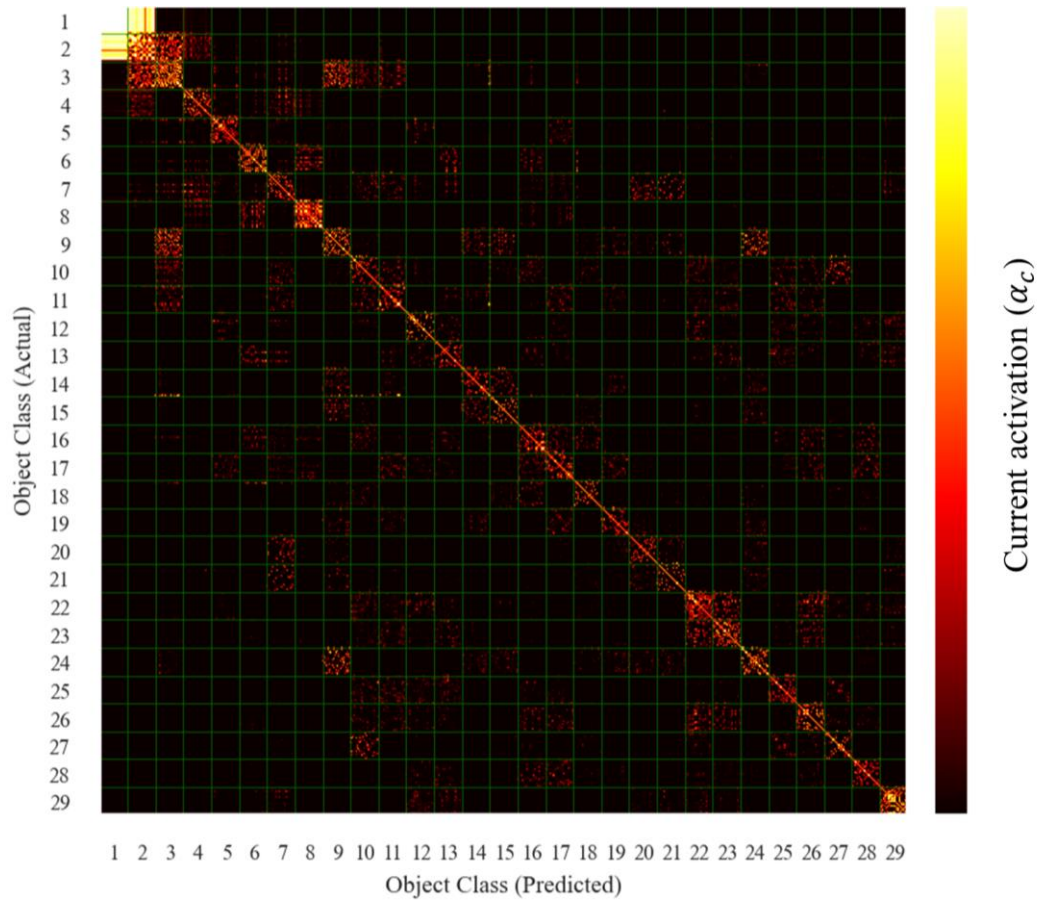


Figure 44. A heatmap showing current activations.

Latent Representation Feature Analysis																	
<b>scale</b>	-2.0	-1.75	-1.5	-1.25	-1.0	-0.75	-0.5	-0.25	0.0	0.25	0.5	0.75	1.0	1.25	1.5	1.75	2.0
<b>rotation</b>	-2.0	-1.75	-1.5	-1.25	-1.0	-0.75	-0.5	-0.25	0.0	0.25	0.5	0.75	1.0	1.25	1.5	1.75	2.0
<b>horizontal translation</b>	-2.0	-1.75	-1.5	-1.25	-1.0	-0.75	-0.5	-0.25	0.0	0.25	0.5	0.75	1.0	1.25	1.5	1.75	2.0
<b>vertical translation</b>	-2.0	-1.75	-1.5	-1.25	-1.0	-0.75	-0.5	-0.25	0.0	0.25	0.5	0.75	1.0	1.25	1.5	1.75	2.0
	<i>extrapolation</i>								<i>interpolation</i>				<i>extrapolation</i>				

Figure 45. Interpolation and extrapolation from latent representations. Numbers above the shapes indicate a step size. The step direction was determined by subtracting the latent representations for the shapes in the blue bounding boxes (for each row).



## Permission Letters

### Permissions for Figure 28

JOHN WILEY AND SONS LICENSE TERMS AND CONDITIONS

Sep 17, 2022

This Agreement between Sean Kugele ("You") and John Wiley and Sons ("John Wiley and Sons") consists of your license details and the terms and conditions provided by John Wiley and Sons and Copyright Clearance Center.

License Number	5391251032460
License date	Sep 17, 2022
Licensed Content Publisher	John Wiley and Sons
Licensed Content Publication	Child Development
Licensed Content Title	Age Differences in Imagery Abilities
Licensed Content Author	Philip F. Daly, Emily J. Goldknopf, Anna M. Barrett, et al
Licensed Content Date	Jun 28, 2008
Licensed Content Volume	61
Licensed Content Issue	4
Licensed Content Pages	16
Type of use	Dissertation/Thesis
Requestor type	University/Academic
Format	Print and electronic
Portion	Figure/table
Number of figures/tables	1
Will you be translating?	No
Title	Embodied, Simulation-Based Cognition: A Hybrid Approach
Institution name	University of Memphis
Expected presentation date	Dec 2022
Portions	Figure 4 on p. 1007

## Permissions for Figure 37

### Perceptual symbol systems



Author: Lawrence W. Barsalou  
Publication: Behavioral and Brain Sciences  
Publisher: Cambridge University Press  
Date: Aug 1, 1999

*Copyright © 1999 Cambridge University Press*

#### License Not Required

Permission is granted at no cost for use of content in a Master's Thesis and/or Doctoral Dissertation. If you intend to distribute or sell your Master's Thesis/Doctoral Dissertation to the general public through print or website publication, please return to the previous page and select 'Republish in a Book/Journal' or 'Post on intranet/password-protected website' to complete your request.

[BACK](#)

[CLOSE](#)