A study in activation: Towards a common lexicon and functional taxonomy in cognitive architectures

S. Kugele

S. Franklin

A Study in Activation: Towards a Common Lexicon and Functional Taxonomy in Cognitive Architectures

Sean Kugele (seankugele@gmail.com) and Stan Franklin (franklin.stan@gmail.com)
Department of Computer Science and Institute for Intelligent Systems, University of Memphis
Memphis, TN 38152 USA

Abstract

Activation has become a pervasive concept in many scientific disciplines, including cognitive and neural modeling, and AI. Unfortunately, its applications and functions are so broad and varied that it is difficult for practitioners to discuss the topic in precise and meaningful ways. This is particularly apparent in cognitive architectures, where a wider breadth of activation's utilities and forms have been explored. To help combat these terminological difficulties, and hopefully facilitate productive discourse and the development of future applications, we introduce (1) a *lexicon* of activation-related concepts, and (2) a *functional taxonomy* that enumerates many activation-related "design patterns" that have appeared in cognitive architectures. We demonstrate our taxonomy by applying it to the LIDA cognitive architecture, which includes one of the most varied and comprehensive adoptions of activation-related functionality.

Keywords: activation, cognitive architectures, LIDA

Introduction

The concept of "activation" has become commonplace in many scientific disciplines. In the neural sciences, it refers to "patterns of neural firing," measured either individually or collectively. In chemistry, it refers to the transition of a molecule to a state with an "increased probability" of a chemical reaction. And, in psychology, it has been used to refer to the "level of arousal or excitation" observed in an individual "as a whole" (Duffy, 1957).

Within the fields of artificial intelligence (AI) and cognitive modeling, activation describes an assortment of quantities and parameters, which have been used to implement a diverse range of functionality. Certainly, activation and "activation functions" have figured prominently in the development of artificial neural networks (ANNs). However, the full breadth of activation's utility and forms has come to fruition in the many cognitive architectures that have embraced, and expanded on, the concept. Unfortunately, these applications and functions are so broad and varied that it is difficult for practitioners to discuss the topic in precise and meaningful ways.

To get a better sense for this diversity, we reviewed seventy-eight cognitive architectures[1] in search of distinct activation-related concepts and themes. Thirty-three of these were found to use some form of activation. While a comprehensive survey of each cognitive architecture's use of activation is beyond the scope of this paper, we believe that we have succeeded in identifying the major concepts and themes. We have distilled these into a *lexicon* of activation-related terminology and a *functional taxonomy* that categorizes each theme with respect to its functional intent (that is, what activation affects, facilitates, or enables within a cognitive architecture). We believe that this is an essential first step towards standardizing notions of activation across cognitive architectures.

Among the cognitive architectures, LIDA (Franklin, et al., 2016) implements one of the most varied and comprehensive adoptions of activation. It uses activation to support nearly every module and process, and one or more activation parameters are associated with most (if not all) of its mental representations. Given this abundance of function and form, LIDA provides a plentiful source of examples, which we use to test our taxonomy's utility, and illustrate its taxonomic themes.

What is Activation?

Arguably, the earliest application of activation-related concepts in AI and cognitive modeling occurred in the context of connectionist models, such as artificial neural networks (ANNs). ANNs are biologically-inspired computational systems composed of "artificial neurons" (also called neural units) that are typically connected in layered architectures. Each neural unit performs a calculation (for example, a weighted sum) over its inputs, and the result is referred to as that neuron's "activation." Activations are then passed through "activation functions" (for example, unit step, sigmoid, or rectified linear) to determine a neuron's output (or response)[2]. These response values can "propagate" to connected neural units, where they are used as inputs to further computations. (This is also referred to as "spreading activation.") ANNs learn "distributed representations" corresponding to the *patterns of activation* induced in the network by various stimuli.

Cognitive architectures have expanded on these connectionist concepts, inventing a host of new mechanisms with their own distinct dynamics. To make sense of this variability, we introduce a basic lexicon of activation-related concepts, and then we review and categorize noteworthy applications of activation within cognitive architectures. This is then used as the catalyst for our

[1] This included most of the cognitive architectures mentioned in (Kotseruba & Tsotsos, 2018), as well as several that were not mentioned.

[2] In the literature, the output of an artificial neural unit is sometimes referred as the unit's "activation"; however, we use the convention that activation refers to the "internal state" resulting from a calculation over its inputs and weights that may be passed as input to an activation function.

"functional taxonomy," which is presented later in the paper.

Concepts and Terminology

Activations are typically implemented as continuous, scalar quantities, or vectors of such quantities (such as, "patterns of activation"). They must have an *activation source* ("how activation is acquired?"), an *activation target* ("what gets activated?"), and a *function(s)* in the system. An implicit or explicit *decay strategy* must also be provided. Optionally, an *activation spreading* mechanism can be implemented that propagates activation between one or more activation targets, and an *activation threshold* can be specified that requires a target's activation be above, or below, a specific value before initiating its associated functionality.

Activation Sources. Anything can be used as a source of activation. The only requirement is that it activates its target *consistently* with respect to some intended purpose. For example, if an activation parameter is intended as a measure of its target's "relevance", "urgency", "salience", "reliability", etc., then the activation source must generate activation in proportion to the target's current compatibility with that measure.

In practice, an *activation source* is often a (mathematical) function specified in terms of other activations or conceptual quantities. An example of this is ACT-R's (Anderson, et al., 2004) formula for determining the activation of a "chunk" (that is, a declarative unit of knowledge):

$$A_i = B_i + \sum_j W_j S_{ji},$$

where A_i is the ith chunk's activation, B_i is its *base-level activation*, W_j is the "attentional weighting" of the jth element in the "current goal," and S_{ji} is the *associative activation* between chunk i and its jth supporting element. In this context, base-level activation and associative activation are activation sources that determine the activation of a chunk. Base-level activation (B_i) has its own activation source, based on the function

$$B_i = \ln\left(\sum_j t_j^{-d}\right),$$

where t_j is the time that has passed since the jth retrieval of chunk i, and d is a fixed parameter that determines the shape of the learning/forgetting curve.

Activation Targets. "Knowledge" representations (semantic, perceptual, procedural, etc.) are the most common *activation targets* in cognitive architectures. However, other data structures, processes, modules, and even entire cognitive systems have been used. The only requirement is that the target is an entity whose function (or identity) can be meaningfully modulated (or determined) by an activation's value.

Decay Strategies. *Explicit* decay strategies are often implemented using (mathematical) functions that are periodically invoked for the purpose of decreasing the value of an activation variable over time. For example, DUAL (Kokinov, 1994) invokes an exponential decay function each time step to decrease the activation of its working memory elements (save for the most active, which is referred to as "the focus"). By contrast, *implicit* strategies often implement decay directly within activation sources; for example, in ACT-R's base-level activation equation (shown earlier), the contribution of the jth chunk retrieval "decays away as a power function (producing the power law of forgetting)" (Anderson, et al., 2004).

Spreading Activation. Many cognitive architectures allow activation to "propagate" between *associated activation targets*. For example, Copycat's (Hofstadter & Mitchell, 1995) long-term memory module, the Slipnet, supports activation spreading between its nodes. These nodes (representing concepts) serve as activation targets, and its links (associations) serve as conduits that allow the passing of activation between them. The "conceptual distance" between nodes, which is based on the activation of its links' labels, determines the amount of activation spread.

A more sophisticated example occurs in the Agent Network Architecture (ANA) (Maes, 1991), which features predecessor, successor, and conflictor links that allow "activation energy" to spread between, and accumulate in, different competence modules (for example, action and belief modules). Predecessor and successor links are *excitatory* (increasing the activations of their associated activation targets) while conflictor links are *inhibitory* (decreasing the activations of their associated activation targets). The magnitude of this increase or decrease is proportional to a (source) competence module's activation.

As a final example, 4CAPS (Just & Varma, 2007) supports activation spreading using weighted *condition-action* production rules that function like weighted links. These productions spread activation iteratively (that is, over multiple "cycles") from task-activated cortical "centers" to their associated declarative elements.

Activation Thresholds. Activation is often functionally "inert" (that is, its associated functionality is not invoked) unless its value crosses above or below an *activation threshold*. Such thresholds control the retrieval of ACT-R's chunks (Anderson, et al., 2004), the execution of ANA's competence modules (Maes, 1991), and the spreading of activation and updating of network weights (associative learning) in LEABRA (O'Reilly, 1996).

Activation thresholds in cognitive architectures can be thought of as generalizations of the binary threshold functions (unit step functions) that appeared in early connectionist networks, such as perceptrons (Rosenblatt, 1958). However, while the function of a perceptron's activation threshold (θ) is limited to modulating its units' "all-or-nothing" output signals ($+1$ if a unit's activation $\geq$

θ; 0 otherwise), the range of functionality that can be modulated in cognitive architectures is seemingly limitless.

Activation's Functions in Cognitive Architectures

Having established a basic vocabulary of activation-related concepts, we now present the major activation-related functional themes that have appeared in cognitive architectures.

Access to Mental Representations. One of the most common uses of activation in cognitive architectures has been to influence global, module-specific, or process-specific access to mental representations. Activation, in this context, can be interpreted as specifying the current, context-specific *relevance* of mental representations, allowing cognitive resources to be focused on representations that matter most at a given moment. For example, in ACT-R, activation controls both the probability and timing of declarative memory (that is, "chunk") retrieval. In Soar (Laird, 2012), activation biases the retrieval of episodic memories. And, in CERA-CRANIUM (Arrabales, Ledezma, & Sanchis, 2009), the priority of percept processing is determined by activation, where those with the lowest activations are not processed at all.

Removal of Mental Representations. Activation has been used to modulate the purging and/or pruning[3] of short-term and long-term mental representations. For example, Soar removes working-memory elements when their activations have decayed below some fixed (removal) threshold (Laird, 2012). These representations are still available in long-term semantic memory for later retrieval, but the system has determined that they are no longer directly relevant to its current task (based on their activations). In other words, these representations are "gone, but not forgotten." We refer to this as "bounded" removal. Representations can also be removed "globally," such as occurs when the base-level activations associated with LIDA's (Franklin, et al., 2016) long-term memory representations (declarative, perceptual, procedural, etc.) decay below a removal threshold. In these cases, the representations are no longer available for use or retrieval. That is, they have been "forgotten."

This functional theme complements activation's use as an *access modulator*, and both uses often appear together in cognitive architectures. Jointly, they can be said to determine the "availability" of mental representations.

Informational Content. Cognitive architectures have used *patterns of activation* as "informational content." These activation patterns must somehow represent the sensory, perceptual, and/or conceptual essences of experiences, introspections, etc. This theme is exemplified in ART (Grossberg, 1999), where patterns of activations are stored as short-term or long-term memory "traces." Shanahan's (2006) brain-based implementation of Global Workspace Theory (Baars, 1988) also makes extensive use of patterns of activation as mental representations.

[3] Pruning refers to the extraction of a mental representation from a data structure (like a tree or associative network) that requires additional structural maintenance (such as the removal of links or "dangling" associations) to purge the targeted item.

Associative Dynamics. Activation has been used to represent the time-varying, context-sensitive, strength of associations between mental representations. Here, activation can be interpreted as associative weights, or modulators of associative weights, whose values are influenced by situational context or prior experiences. ACT-R's "associative activations" are one example of this. Another example occurs in Copycat's Slipnet, where the conceptual distances between its nodes are based on the activations of its links' "labels." Concepts (that is, nodes) that are closer in conceptual distance (that is, have higher link label activations) are more likely to "slip" into one another, and be treated as analogous concepts.

System Dynamics. Activation can locally or globally modulate "how" cognitive operations are performed. In this context, activation could be viewed as representing dynamic dispositions, temperaments, or moods, and the notion of "activation as arousal" (see (Duffy, 1957)) is consistent with this theme. An example of this from a cognitive architecture is Copycat's "temperature," which is described by Hofstadter and Mitchell (1995) as a variable that "monitors the stage of processing, and helps to convert the system from its initial largely bottom-up, open-minded mode to a largely top-down, closed-minded one."

Measures of Intensity or Degree. Activation often represents a graded measure of some quantity with a clear semantic interpretation (for example, "reliability") that fluctuates in intensity over time. Due to the conceptual interpretability of these measures, they often serve as activation sources that modulate other cognitive functions. An example of activation as a *measure of intensity* appears in Leabra, where network activations represent "graded (continuous) states of truth-value" that estimate "the degree to which [a hypothesis] is believed to be true by the network" (O'Reilly, 1996). Another example occurs in LIDA, where the activations associated with LIDA's feeling nodes quantify an agent's current "liking" or "disliking" of a stimulus.

Process Scheduling. Activation has been used to determine or influence the execution of events, tasks, processes, and modules. The CopyCat architecture contains a module called the Coderack that serves as a pool of "codelets[4]." Each codelet is associated with an "urgency" value that is a function of the activation patterns in the Slipnet. These are used by the Coderack to determine the *probability* that a particular codelet will be selected for execution. Activation,

[4] Codelets are processes that function as simple agents with the ability to find, create, or destroy structures in Copycat's Workspace.

in this context, can be viewed as *influencing system dynamics* through the immediate or future execution of some cognitive process. Another example occurs in DiPRA (Pezzulo, 2009), which contains an "energy pool" from which its modules receive activation at the beginning of each execution cycle. Since activation is required for module execution, if the energy pool is depleted in a given cycle, then one or more modules may have to wait until a later cycle to execute.

A Functional Taxonomy

In this section, we present our "functional taxonomy" of activation-related parameters/variables based on their uses in cognitive architectures (see Figure 1). At the highest level of our taxonomy, we divide activation-based functionality into three major themes: "representational," "system dynamics," and "measures of intensity or degree."

The *representational branch* is sub-divided into "associative dynamics," "availability," and "informational content." *Associative dynamics* includes "activation spreading" and the activation-based modulation of representational associations (for example, Copycat's conceptual distances). *Availability* covers the global (that is, system-wide) and bounded (that is, process or module-level) access and removal of representations. *Global*, in this context, could correspond to "forgetting" from long-term memory, and *bounded* to the eviction of representations from short-term memory. Lastly, *informational content* covers use-cases like ART's memory traces.

The *system dynamics* branch is sub-divided into "modulatory" and "scheduling" functions. *Modulatory* functions include system-wide, module-specific, or process-specific activation parameters that influence "how" operations are executed. This includes Copycat's temperature, and the notion of "arousal" from the psychological literature. *Scheduling* refers to the *deterministic* or *probabilistic* initiation of events or processes, based on activation, resulting in short-lived or long-lasting changes to a system's dynamics. LIDA's "triggers"[5] are examples of deterministic scheduling. Copycat's codelet "urgency" values are examples of probabilistic scheduling.

The *measures of intensity or degree* branch is intended to cover all activation parameters that serve to label an activation target as possessing some degree of an unambiguously defined property. This property should have a clear semantic interpretation. The magnitude of the activation indicates "to what extent" that target possesses the property. This covers, for example, Leabra's use of activation as a measure of the "truthiness" of a hypothesis.

Our Taxonomy Illustrated in LIDA

Activation is ubiquitous in LIDA, with activation-related variables and parameters supporting most (if not all) of its

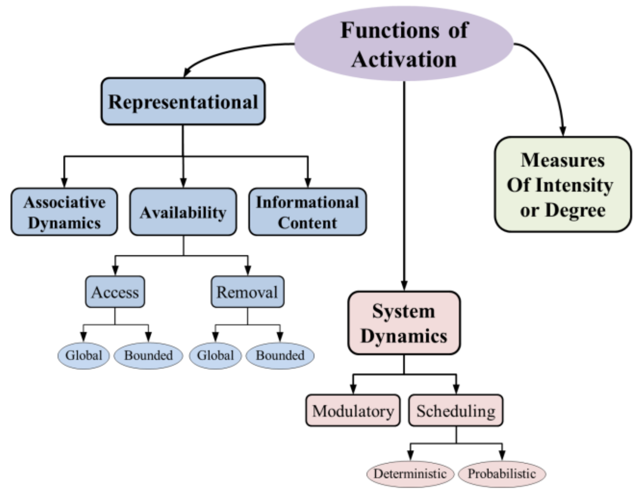

Figure 1: A "functional taxonomy" of activation. Each element specifies a category of functions that an activation variable or parameter could support.

modules, processes, and data structures. All the major themes in our functional taxonomy are present in LIDA; therefore, in this section, we use LIDA to illustrate our functional taxonomy. But first, we introduce LIDA and its activation-related concepts, so that the reader is prepared for the demonstrations that follow.

What is LIDA?

Learning Intelligent Decision[6] Agent (LIDA) (Franklin, et al., 2016) is a biologically-inspired cognitive architecture that implements, and fleshes out, significant portions of the Global Workspace Theory (GWT) of consciousness (Baars, 1988), as well as many other psychological theories (for example, Baddeley & Hitch, 1974; Barsalou, 1999; Conway, 2001; Ericsson, 1995). LIDA contains numerous short-term memory (STM) and long-term memory (LTM) modules, and special purpose processors called codelets[7]. These are depicted in Figure 2, and their functions and common acronyms are summarized in Table 1.

Cognition occurs in LIDA over a continual series of potentially overlapping "cognitive cycles," which correspond to the "action-perception cycle" referred to by many psychologists and neuroscientists (Fuster, 2004; Neisser, 1976). Each cognitive cycle is conceptually divided into "perception and understanding," attention, and "action and learning" phases. Higher-order cognitive processes such as planning, deliberation, and problem solving typically require many cognitive cycles. See (Franklin, et al., 2016) for more details.

LIDA's Activation Concepts

LIDA has historically classified its activation parameters as either "base-level activations," "current activations," or

[5] LIDA's triggers are explained in more detail later in the section entitled "Triggers."

[6] For historical reasons, this word was previously "distribution". It was later changed.

[7] This terminology was inspired by Copycat's codelets.

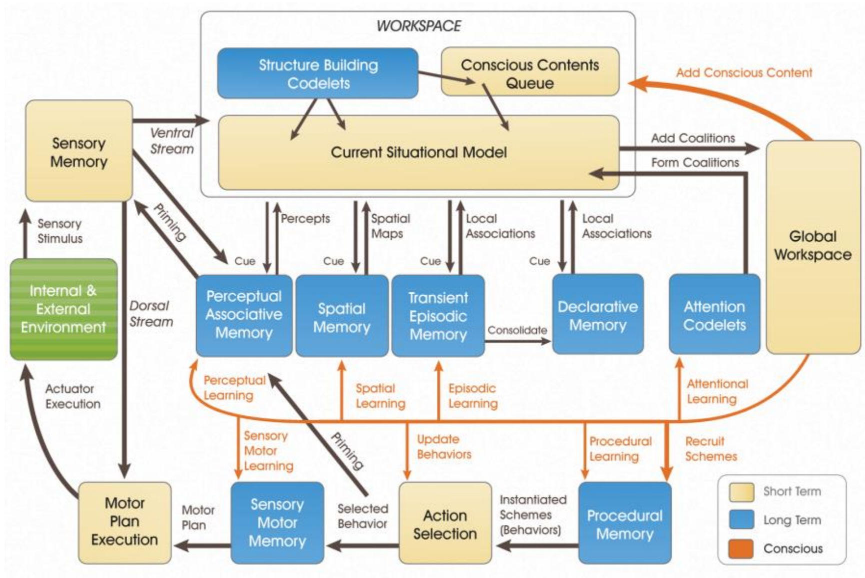

Figure 2: The LIDA cognitive cycle diagram.

simply "activations" (or "total activations"). *Base-level activation*[8] is used to describe activations with relatively slow decay rates that have activation sources based on content in the global "conscious" broadcast. These activations support "selectionist learning" (Edelman, 1987), and are the basis for the removal (forgetting) of long-term memory representations and processes. *Current activation* refers to parameters with relatively rapid decay rates that (generally) reflect transitory, module-specific notions of current "relevance." And, *activation* (or *total activation*) is used to describe all other activation parameters. Many of these general-purpose activation parameters use base-level and current activations as activation sources; however, this is not always the case.

Taxonomic Examples in LIDA

In this section, we will illustrate our taxonomy using activation-related examples from LIDA. These examples do not represent an exhaustive account of LIDA's activations; however, they should be sufficient to give the reader a taste of LIDA's major activation themes, and build an intuition for how our taxonomy could be applied in practice. Following each sub-section, we summarize the taxonomic themes that were covered.

Low-Level Feature Detectors and SM Representations. Modality-specific, low-level features detectors in Sensory Memory (SM) are activated in response to incoming sensory stimuli (from an agent's sensors). The *patterns of activation* generated in these feature detectors serve as sensory representations, corresponding to the incoming stimuli, that can be later incorporated into knowledge representations in the Workspace and long-term memory modules.

Taxonomic Themes Illustrated:

1. **Representational → Informational Content**
   [*sensory representations as "activation patterns"*]

The Activation and Instantiation of Percepts. SM uses its sensory representations to activate (that is, update the "current activations" of) feature detectors in Perceptual Associative Memory (PAM). Activation then spreads, over "activation links," to linked PAM nodes. A PAM node's activation is based on the sum of its base-level and current activations. PAM nodes with activations greater than a fixed threshold are instantiated into the Current Situational Model (CSM) as percepts, making them available to codelets and cueing processes.

Taxonomic Themes Illustrated:

1. **Representational → Associative Dynamics**
   [*spreading activation in PAM*]
2. **Representational → Availability → Access → Bounded**
   [*percepts to CSM*]

Attention Codelets and Coalitions. Preconscious content in the CSM (including percepts and other LTM representations) retain their activations after instantiation (though they subsequently decay). Some attention codelets (ACs) (for example, the "default attention codelet" described in (Franklin, et al., 2016)) use the activation associated with preconscious content to determine their "level of interest" in those representations. When an AC is "sufficiently interested" in a representation, it will take it to a coalition forming process, which may create a *coalition* containing that content. A coalition's activation is based on the activation of its content, and the base-level activation of the AC advocating for that content (among other things).

[8] LIDA's base-level activation is roughly (conceptually) analogous to ACT-R's 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.

Table 1: Descriptions of LIDA's short-term memory (STM) and long-term memory (LTM) modules, and codelets.

Module / Process	Description
ACTION SELECTION (AS)	STM module supporting the selection of behaviors for execution by the SMS.
ATTENTION CODELETS (ACs)	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.
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 *coalitions*, 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 *percepts* after being activated by incoming sensory content (or cueing).
PROCEDURAL MEMORY (PM)	LTM module containing representations called *schemes* 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) *behaviors*.
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.

Coalitions compete in a winner-take-all competition in the GW, *based entirely on the coalitions' activations*. The winning coalition's content is broadcast, making it globally accessible to all modules and processes.

Taxonomic Themes Illustrated:

1. **Representational → Availability → Access → Bounded** [*CSM to coalition forming process via default AC*]
2. **Representational → Availability → Access → Global** [*GW to global broadcast*]

Triggers. Competitions are *triggered* in the GW when a *single coalition* has an activation greater than an activation threshold, or a *set of coalitions* has activations *collectively* greater than a (different) threshold. LIDA's Action Selection module also features triggers that initiate competitions among its *behaviors* based on their activations. Since activation, in these cases, influences the rate at which conscious broadcasts occur, and actions are selected for execution, they are great examples of how activation can directly, and deterministically, influence a system's *dynamics* through event scheduling.

Taxonomic Themes Illustrated:

1. **System Dynamics → Scheduling → Deterministic** [*triggers*]

Strength of Global Broadcasts and Attentional Blinks. The *strength* of a global broadcast is determined by the magnitude of the winning coalition's activation, which is used to modulate base-level activation updates in LIDA's LTM modules (that is, selectionist learning), and update other activations in STM modules. If a broadcast's strength is "extremely" high, it can induce an "attentional blink" (Madl & Franklin, 2012); that is, a brief "refractory period" that affects all ACs, from which it gradually recovers. During the refractory period, coalitions receive less activation when they are formed; therefore, conscious broadcasts are more likely to be triggered based on the elapsed time since the last broadcast, than the coalitions' activations.

Taxonomic Themes Illustrated:

1. **Measures of Intensity or Degree**
   [*strength of conscious broadcast*]
2. **System Dynamics → Modulatory**
   [*attentional refractory period*]

Affective Valence and Feelings. LIDA's motivational system (McCall, Franklin, Faghihi, Snaider, & Kugele, 2020) is grounded in "feeling nodes"—PAM nodes with *affective valence*. Affective valence is a form of activation that quantifies notions of liking or disliking with respect to drives (hunger, thirst, etc.), or other interpretative aspects (sweetness, warmth, etc.) of real or imagined events.

Taxonomic Themes Illustrated:

1. **Measures of Intensity or Degree**
   [*feelings*]

Closing Remarks

In this paper, we've presented a lexicon of activation-related concepts, and a functional taxonomy that characterizes how activation has been historically applied in cognitive architectures. While we have made our best effort at gathering the major concepts and themes, it's likely that others remain. Similarly, the validity and usefulness of our taxonomy requires additional testing. Nevertheless, we hope that our efforts towards a common vocabulary will inspire activation-related discussions, and lead to a greater understanding of the concept as a whole.

Acknowledgements

We would like to thank Dr. Pulin Agrawal for his thoughts and guidance at the onset of this endeavor.

References

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.

Arrabales, R., Ledezma, A., & Sanchis, A. (2009). CERA-CRANIUM: A test bed for machine consciousness research. *International Workshop on Machine Consciousness.* Hong Kong.

Baars, B. J. (1988). *A Cognitive Theory of Consciousness.* Cambridge University Press.

Baddeley, A. D., & Hitch, G. J. (1974). Working memory. In G. A. Bower (Ed.), *The Psychology of Learning and Motivation* (pp. 47-89). New York: Academic Press.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and brain sciences, 22*(4), 577-660.

Conway, M. A. (2001). Sensory–perceptual episodic memory and its context: autobiographical memory. *Philos. Trans. R. Soc. Lond B., 356*, 1375–1384.

Duffy, E. (1957). The psychological significance of the concept of "arousal" or "activation". *Psychological Review*, 265-275.

Edelman, G. M. (1987). *Neural Darwinism: The theory of neuronal group selection.* Basic books.

Ericsson, K. A. (1995). Long-term working memory. *Psychological Review*, 211–245.

Franklin, S., Madl, T., Strain, S., Faghihi, U., Dong, D., Kugele, S., . . . Chen, S. (2016). A LIDA cognitive model tutorial. *Biologically Inspired Cognitive Architectures*(16), 105-130.

Fuster, J. M. (2004). Upper processing stages of the perception–action cycle. *Trends in cognitive sciences, 8*(4), 143-145.

Grossberg, S. (1999). The link between brain learning, attention, and consciousness. *Consciousness and cognition, 8*(1), 1-44.

Hofstadter, D. R., & Mitchell, M. (1995). The copycat project: A model of mental fluidity and analogy-making. In K. J. Holyoak, & J. Barnden (Eds.), *Advances in connectionist and neural computation theory, Vol. 2: logical connections* (pp. 205-267). Norwood N.J.: Ablex.

Just, M. A., & Varma, S. (2007). The organization of thinking: What functional brain imaging reveals about the neuroarchitecture of complex cognition. *Cognitive, Affective, & Behavioral Neuroscience, 7*(3), 153-191.

Kokinov, B. N. (1994). The DUAL Cognitive Architecture: A Hybrid Multi-Agent Approach. In *11th European Conference on Artificial Intelligence* (pp. 203-207).

Kotseruba, I., & Tsotsos, J. K. (2018). 40 years of cognitive architectures: core cognitive abilities and practical applications. *Artificial Intelligence Review*, 1-78.

Laird, J. E. (2012). *The Soar cognitive architecture.* MIT press.

Madl, T., & Franklin, S. (2012). A LIDA-based model of the attentional blink. In *Proceedings of the 11th International Conference on Cognitive Modelling (ICCM 2012)* (pp. 283-288).

Maes, P. (1991). The agent network architecture (ANA). *Acm sigart bulletin, 2*(4), 115-120.

McCall, R. J., Franklin, S., Faghihi, U., Snaider, J., & Kugele, S. (2020). Artificial Motivation for Cognitive Software Agents. *Journal of Artificial General Intelligence, 11*(1), 38-69.

Neisser, U. (1976). *Cognition and reality: Principles and implications of cognitive psychology.* San Francisco: W.H. Freeman.

O'Reilly, R. C. (1996). *The Leabra model of neural interactions and learning in the neocortex.* Doctoral dissertation, PhD thesis, Carnegie Mellon University, Pittsburgh, PA.

Pezzulo, G. (2009). DiPRA: a layered agent architecture which integrates practical reasoning and sensorimotor schemas. *Connection Science, 21*(4), 297-326.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review, 65*(6), 386-408.

Shanahan, M. (2006). A cognitive architecture that combines internal simulation with a global workspace. *Consciousness and cognition, 15*(2), 433-449.