Standard model of the mind and perceptual control theory – A theoretical comparison between two layouts for cognitive architectures

Anni-Emilia Huuhtanen

MSc thesis
UNIVERSITY OF HELSINKI
Department of Computer Science

Helsinki, May 1, 2019

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet — Faculty	Laitos — Institution — Department
Faculty of Science	Department of Computer Science

Tekijä — Författare — Author
Anni-Emilia Huuhtanen

Työn nimi — Arbetets titel — Title
Standard model of the mind and perceptual control theory – A theoretical comparison between two layouts for cognitive architectures

Oppiaine — Läroämne — Subject
Computer Science

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
MSc thesis	May 1, 2019	46

Tiivistelmä — Referat — Abstract

Research on artificial intelligence (AI) has often focused on techniques for implementing specialized aspects of intelligence without regard for the full picture of intelligence and cognition. However, a need for more general, human-like AI systems has also been recognized. In this thesis, our aim was to study potential starting points for a comprehensive, general and truly human-like cognitive architecture. We performed a comparative theoretical analysis on two existing layouts, the standard model of the mind and perceptual control theory (PCT), based on functional criteria gathered from literature. While our results indicate that the PCT model is more comprehensive than the standard model on a theoretical level, finding out the true benefits and challenges of the models and their suitability as foundations for human-like AI systems requires practical evaluation. Additional research is needed to fill current gaps in both models, create their computational implementations, and design practical evaluation methods suitable for them.

ACM Computing Classification System (CCS):

Computing methodologies → **Cognitive science**
Computer systems organization → Self-organizing autonomic computing

Avainsanat — Nyckelord — Keywords
artificial intelligence, cognitive architecture, standard model of the mind, perceptual control theory

Säilytyspaikka — Förvaringsställe — Where deposited

Muita tietoja — Övriga uppgifter — Additional information

Contents

1 Introduction

Contemporary research on artificial intelligence (AI) tends to be more analytic than synthetic [16]. Analytic AI research focuses on a certain aspect of intelligence such as perception, learning or problem-solving in detail and often in isolation with other aspects. For example, in recent times, AI systems based on statistical techniques such as deep neural networks (DNNs) capable of recognizing objects, processing speech and playing video games have received a lot of research efforts and public attention. In contrast to analytic AI research, synthetic AI research focuses on the interaction between several aspects of intelligence with the goal of more generally intelligent systems that are able to perform various tasks in complex environments.

Several problems are associated with current, mostly deep learning-based AI systems that originate from analytic research [15, 18, 7, 25, 19]. First, they tend to be data-hungry, requiring a lot of training examples in order to learn and being unable to learn abstract rules, causal relationships and concepts. Second, they usually have a narrow area of competence and may even make rather surprising errors in their specialized domain. For example, an object recognition system trained on images of guns does not recognize dogs, and in some cases it can falsely classify a turtle as a gun [8]. Third, it is not always easy for humans to understand how and why these systems work the way they do (the *black box* problem), complicating interactions and trust between them and humans. Humans cannot apply their existing mental models about the human mind (*theory of mind*) to them in order to predict and reason about their motivations and behavior: their inner workings or ways of processing information (mainly based on statistics) are somewhat different, and they do not share the same language, gestures and other means of communication that are familiar to us. This is not necessarily a problem if the systems are designed as *cognitive prostheses*, i.e. to replace human capabilities and require zero collaboration with humans. However, when their purpose is to complement and closely collaborate with humans (*cognitive orthotics*), it is especially important that the system is capable of human-level interaction and promotes trust.

To solve these problems, the synthetic approach to AI end especially taking inspiration from the human mind in designing general AI systems is promising

- after all, humans currently are the best example of general intelligence [15, 19]. For instance, humans are competent in several complex domains, can often learn from just a few examples, and excel at learning abstract, causal and conceptual knowledge. In addition, psychological similarity between humans and machines could improve understanding, trust and collaboration between them [31, 7]. Insights on human intelligence (or *cognition*) can be drawn from many fields such as cognitive science, neuroscience and psychology. For example, *cognitive architectures* studied in cognitive science are unified models of the fixed structure and processes of the human mind similar to grand theories in other fields such as physics, and they usually involve both an abstract model and a low-level computational implementation [24, 13]. In the implementation, cognitive structures and processes such as perception, attention, learning, reasoning and language are mimicked by computational methods such as neural and semantic networks, decision trees, description logic, and so on, all in a unified framework. This computational formalization not only helps cognitive scientists develop and validate cognitive theories but may also provide a platform for general AI systems. The recent *standard model of the mind* represents an attempt to reach consensus on cognition and is intended as a general abstract layout for cognitive architectures [14].

Another general model of cognition comes from an old, less-known psychological theory called *perceptual control theory* (PCT) that is based on control theory [20]. Control theory is a mathematical theory that models a process of maintaining variables such as temperature or speed in certain pre-determined goal states amidst changing conditions [29]. PCT is based on the idea that all human behavior, too, can be seen as control of perceptual variables. According to PCT, cognition consists of a hierarchy of controller units that constantly compare input perceptions to goal perceptions and attempt to minimize errors between them. PCT says that it is perceptual input that is under control, not behavioral output. This view is in contrast with the sense-think-act process underlying traditional cognitive architectures and the standard model that say output of the system is some function of input.

In this thesis, our purpose was to find out how promising the standard model and the PCT model are as high-level layouts for comprehensive, human-like cognitive architectures and thus possible paths to general AI. We present a

comparative theoretical analysis of the standard model and the PCT model based on how they address certain functional criteria gathered from literature on human-like cognition. Our precise research questions were as follows.

1. How does the standard model on a theoretical level address the functions required from human-like cognition?
2. How does the PCT model on a theoretical level address the functions required from human-like cognition?

The thesis is organized as follows: In Background, we explain the concepts of cognition and cognitive architecture in more detail and offer a brief introduction to the standard model and the PCT model. In Methods, we explain why we chose to perform a comparative analysis on the specific models and introduce the functional criteria used in the comparison. In Results, we present the results of our comparative analysis. Finally, we end with a discussion of our results and conclude.

2 Background

In this chapter, we introduce the relevant concepts of the thesis. First, we look at how the concept of cognition is defined in different paradigms of cognitive science. Second, we explain the concept of a cognitive architecture in cognitive science. Finally, we introduce the standard model and the PCT model.

2.1 Cognition

A basic tenet in the field of cognitive science is that cognition is some form of *information-processing* [10, 4]. Cognitive science can be divided into several paradigms or schools of thought that approach this differently. Usually, a division has been done between three major approaches: *classical*, *connectivist* and *embodied*. These approaches are summarized in Table 1.

The *classical* approach likens the human mind to a digital computer and interprets information-processing mainly as rule-based symbol manipulation [4]. The focus of this approach is problem-solving in simple and well-defined domains such as games where solutions can be found by searching through a problem space. Cognitive systems studied by the approach mostly include higher-level systems such as reasoning, planning, problem-solving and language.

The *connectivist* approach sees the mind not as a serial, logical digital computer but as a parallel sub-symbolic processor of statistical patterns [4]. It models the mind with artificial neural networks (ANNs) inspired by real neural networks of the human brain. ANNs are able to learn from example and usually deal better with uncertainty than rule-based symbolic processors. Issues pertaining to the "lower" levels of cognition such as perception and motor action have traditionally been the focus of connectivists.

The *embodied* approach emphasizes the role of the whole body in cognitive processing and direct connections between sensing and acting in contrast to the traditional sense-think-act process of classicists and connectivists in which some kind of processing of mental representations (thinking) usually occurs between sensing and action [4]. According to the extended mind hypothesis

Table 1: Different schools of thought in cognitive science.

Approach	**Manner of information-processing**	**Nature of cognitive architecture**	**Modeled aspects of cognition**
Classical	Centralized, rule-based, logical, serial, symbolic	Computer-like: structure and process separate	Abstract reasoning, problem-solving, planning
Connectivist	Decentralized, pattern-based, statistical, parallel, dynamic	Brain-like: structure and process together	Perception, action, low-level reasoning and planning
Embodied	Decentralized, embodied, direct, reactive, dynamic	?	Perception, action

of the embodied approach, the mind is not separated from the environment. A radical embodied view says that interaction between cognition and the environment via the body is so intertwined and real-time that there is no need to form and process any internal knowledge representations about the environment. It is thought that the mind does not necessarily need complex internal models of the world because the world already is its own model. The embodied approach is still relatively new and lacks established models and theories.

2.2 Cognitive architectures

In 1973, Allen Newell pointed out the problem with psychological micro-theories that attempt to explain only specific parts of cognition, prompting a search for a unified theory of cognition - a *cognitive architecture* [24]. Cognitive architectures are abstract models of the fixed structure and processes of the human mind in cognitive science [13, 43, 16]. The term *fixed* means unchangeable over time and between situations. Cognitive architectures attempt to model different cognitive systems or functions and their relationships to each other in a unified manner without focusing solely on a certain singular function. Cognitive functions include, for example, perception, attention, learning, reasoning, problem-solving, planning, language, memory, emotions, motivation, personality, and imagination.

Usually, a cognitive architecture is expressed both as an abstract model as well as an implementation of that model in the form of a software framework,

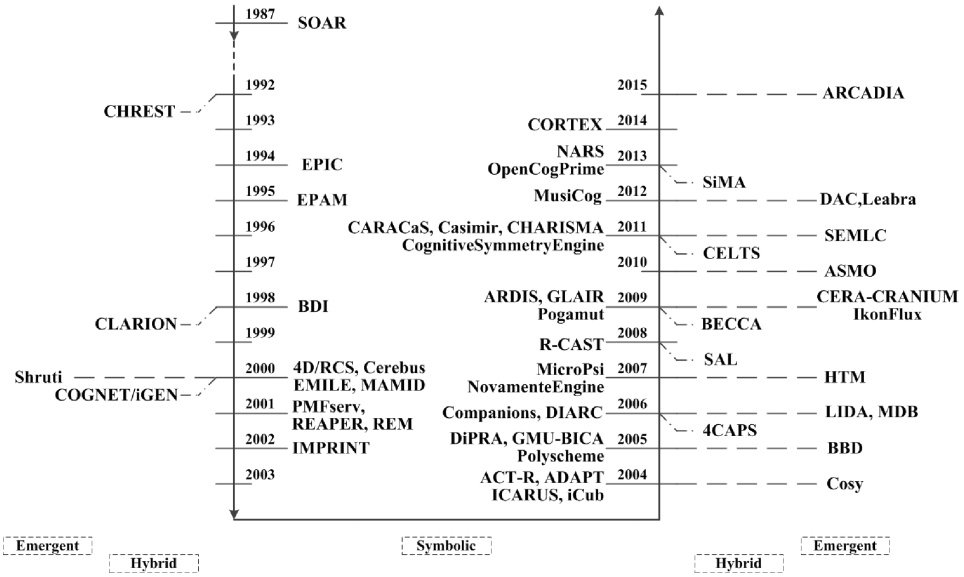

Figure 1: Timeline including a small subset of cognitive architectures [43].

and this framework can then be used to build an instance of artificial cognition for a particular task or purpose [13, 43, 16]. Cognitive architectures differ in terms of how much functionality is included in the framework and how much has to be programmed by the user of the architecture. For example, a cognitive architecture may specify a process and structure for handling certain type of knowledge but the knowledge content itself must be supplied by the user. On the other hand, some cognitive architectures are able to learn the required knowledge from experience without needing as much preprogramming. Many cognitive architectures are open source projects whose contributors include researchers and software engineers from multiple disciplines.

Currently, several hundred cognitive architectures exist [13]. A subset of them on a timeline is shown in Figure 1. Some architectures, for example Soar and ACT-R, have been in development for decades and established themselves as notable candidate theories of cognition. Cognitive architectures differ in terms of their underlying assumptions about how the human mind works, motivation, inspiration, implementation technology, and so on [43]. Certain cognitive functions such as problem-solving, planning and memory have received more attention than other functions such as emotions, motivation, personality and social intelligence.

Cognitive architectures can be categorized as *symbolic*, *hybrid* or *emergent* based on their patterns for processing knowledge [43]. *Symbolic architectures* originate from the classical school of cognitive science and tend to utilize techniques of the GOFAI (Good Old-Fashioned Artificial Intelligence) paradigm of AI. Knowledge is represented as symbols and its processing occurs through formal logic. These architectures may not include as many capabilities for processing low-level information such as sensations or executing motor actions - instead, the focus tends to be on high-level, internal processing. They are especially suitable when deterministic behavior is desired although a few architectures use fuzzy instead of classic logic. Symbolic architectures are less suitable for recognizing patterns in input information and dealing with uncertain input. Examples of these kind of architectures include Soar, ACT-R and ICARUS. *Emergent architectures* originate from the connectivist school of cognitive science and are commonly utilized in behavior-based robotics. They often involve a hierarchical structure that is inspired by the brain. In this hierarchy, the lower levels may process and store sensory information in neural networks or by other statistical techniques (also known as bottom-up processing) while the higher levels control attention and select actions based on the output of the lower levels. Emergent architectures may not include higher-level capabilities such as problem-solving and planning - instead, they focus mainly on the interaction between attention, perception and motor action. Emergent architectures are more capable of dealing with uncertainty and recognizing statistical patterns in input information. Examples of these architectures include ARCADIA, LIDA and BBD. *Hybrid architectures* tend to process both symbolic and emergent knowledge in a hierarchical structure in which sub-symbolic bottom-up knowledge is combined with symbolic top-down knowledge. This seems to most closely follow our intuition of how the human mind works: more symbolic rule-based processing (planning, problem-solving, memory) is combined with numeric pattern-based processing (perception). Examples of hybrid architectures include CLARION, CREST and 4CAPS. In addition to these three classes - symbolic, emergent and hybrid - cognitive architectures can be classified into *embodied architectures* that model the physical body and its role in cognitive processing and *developmental architectures* that take inspiration from human cognitive development.

The existence of symbolic, emergent and hybrid cognitive architectures reflects

the fact that cognitive activity can be modeled at multiple timescales or levels of abstraction [24]. Cognition ranges from fast, concrete and primitive neural processing such as perception to slow, abstract and high-level processing such as language and social cognition. In between, there are cognitive and rational levels that are constrained by the lower neural level and offer a foundation to the higher levels. Emergent architectures mainly model the neural level, symbolic architectures mainly model levels above the neural level called levels of deliberate action, and hybrid architectures model all levels. In some of those cognitive architectures that model levels of deliberate action, higher-level processing is assumed to emerge from the knowledge and skills (memory content) acquired by an architecture through sequential cognitive cycles while in others, higher-level processing has its own dedicated modules.

The computational implementations of cognitive architectures have two benefits [13, 43, 16]. First, they help cognitive scientists validate and further develop their cognitive theories. Various psychological experiments have been successfully replicated with cognitive architectures for validation purposes. Similar performance between a cognitive architecture and a human in a particular experiment gives indication that the architecture accurately models some aspect of human cognition. Second, cognitive architectures are useful platforms for those in the field of AI who are interested in building software applications that exhibit general artificial intelligence. The goal of general AI is to build artificial systems that can function in complex and variable environments and tasks which is a typical human capability requiring interaction between several cognitive functions. Cognitive architectures can be used to build software agents that function in virtual environments or in the real world via a physical robot body, and they have been applied in several ways in practice. One application category is practical tasks, for example navigating and collecting items in unknown environments, tutoring, cleaning, and making medical assessments. Other application categories include systems that are able to collaborate with humans, systems for natural language processing such as syntactic and semantic parsing, classification and pattern recognition, standalone machine vision tasks, video and board games, and virtual agents.

There are a few problems with many existing cognitive architectures [13,

32, 16]. First, they tend to have limited ability to learn through experience and generalize knowledge across tasks. Learning is often focused on physical or mental actions instead of concepts, categories, their relations and other declarative knowledge, and the architectures are unable to truly *understand* and make many different interpretations about often partial knowledge. It is often necessary to preprogram knowledge required in different tasks into cognitive architectures, and an architecture that has been preprogrammed to perform one task usually cannot perform another task. This requirement for ad hoc preprogrammed knowledge has also made it considerably difficult to model those cognitive systems that require lots of prerequisite knowledge such as social intelligence and language. Second, while many cognitive architectures have been shown to excel in different tasks and some psychological experiments have been successfully replicated with them, each task and experiment has usually tested a narrow set of cognitive functions in a somewhat artificial laboratory setting. In actuality, the architectures are often deficient in certain cognitive functions, especially social cognition such as verbal and non-verbal communication, emotions, metacognition, personality, creative thinking, perception, and episodic (autobiographic) memory. All in all, most current cognitive architectures do not seem to be unified theories of cognition (as intended by Newell) as much as they are collections of micro-theories regarding specific functions of cognition.

2.3 Standard model of the mind

Recently, the amount and variability of cognitive architectures provoked a search for a common consensus regarding the mind across neuroscience, cognitive science, robotics and artificial intelligence, resulting in the proposal of a *standard model of the mind* (Figure 2) [14]. Its purpose is to provide a general abstract layout for cognitive architectures, and it is based on three popular cognitive architectures: Soar, ACT-R and Sigma. There is a plan to incorporate knowledge from a larger set of cognitive architectures in the future as new agreements on matters can be reached.

The standard model is modular, consisting of several components that process information independently and share it with each other [14]. The components of the standard model include perception, motor, declarative

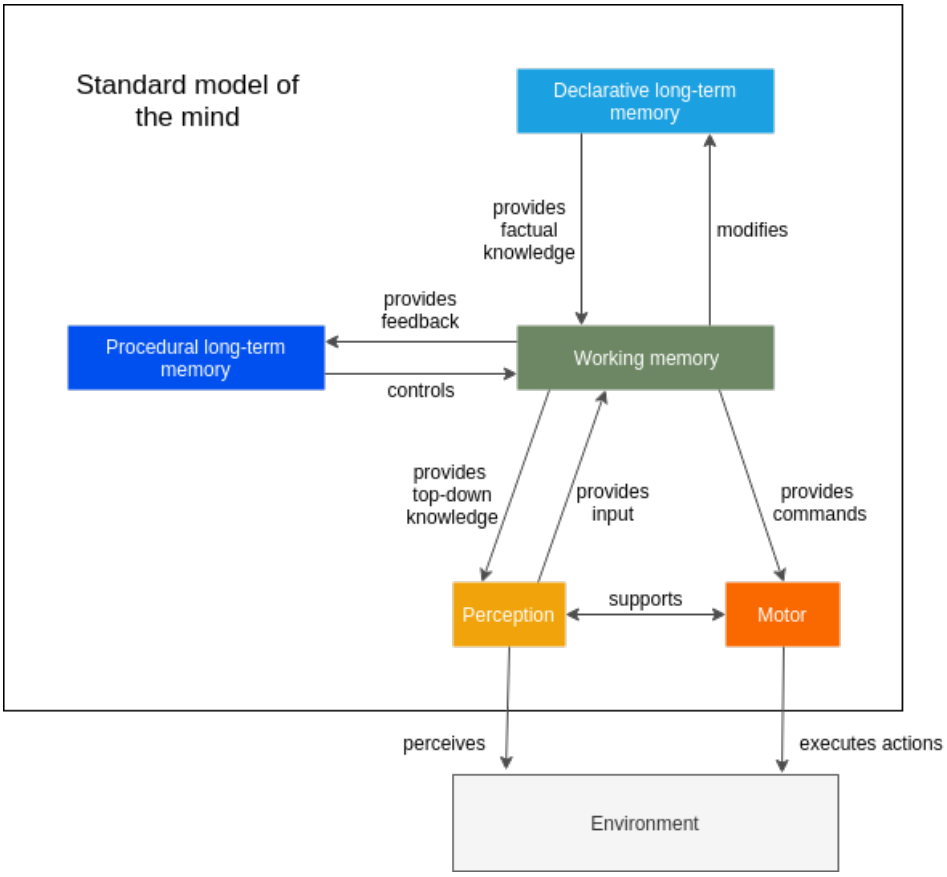

Figure 2: The standard model of the mind (adapted from [14]).

long-term memory, procedural long-term memory, and short-term memory (also called working memory). Each of them can consist of smaller sub-modules, for example declarative long-term memory can be further divided into semantic and episodic memories if desired. Information processing occurs in cognitive cycles of around 50 ms each, i.e. at the level of deliberate action. Complex cognitive phenomena arise from sequences of cognitive cycles. In each cycle, procedural memory selects a single deliberate action that causes modifications in the working memory. A single modification can involve, for instance, a step forward in a logical reasoning process, an initiation of a motor action, or a fetch from long-term memory. Although cognitive cycles run serially, parallel processing of information is possible inside and between components. The standard model is a hybrid model: the memory components process symbolic knowledge combined with numerical

metadata while the perception and motor components process sub-symbolic knowledge.

2.4 Perceptual control theory

Control theory is a mathematical theory that models a process of maintaining variables such as temperature or speed in certain pre-determined goal states amidst changing conditions [29]. In engineering, the theory has been applied to building different self-adaptive systems that must exhibit autonomous and goal-oriented behavior, for example thermostats and cruise control systems in cars. In many of these systems, self-adaptation must occur according to precise requirements and despite restrictions such as time or memory, and with control theory, formal guarantees for self-adaptation can be provided. Control theory has traditionally been used to implement self-adaptivity at the level of hardware, for example CPU and memory. Lately, it has also been applied to software adaptation.

Control theory is also the basis of *perceptual control theory* (PCT) in psychology [20]. PCT is a general psychological theory based on the idea that all human behavior, too, can be seen as control of perceptual variables. According to PCT, it is perception that is under control, not behavior. Behavior is simply varied in order to achieve desired perceptions (*goals*). PCT says, for example, that when a person attempts to catch a ball coming towards them, it seems unlikely that they do complex calculations to figure out where the ball is going to land in order to move to that place. Instead, they simply move in a manner that keeps the perception of the ball a certain way. There is no need for complex and accurate models about the world nor predictions or planning based on them - if planning occurs, it is not about what actions to take in order to achieve goals but about what are the desired perceptions. PCT argues against the more traditional view of cognition as a sense-think-act process that suggests output is some function of input; that perception (input) controls or causes behavior (output) in a linear way. In PCT, the term *perception* refers not to internal or external sensory signals themselves but to the whole set of events that happens in the brain as the signals travel from sensory modalities to the top levels of the brain. These events may include the conceptualization of the signals but not necessarily

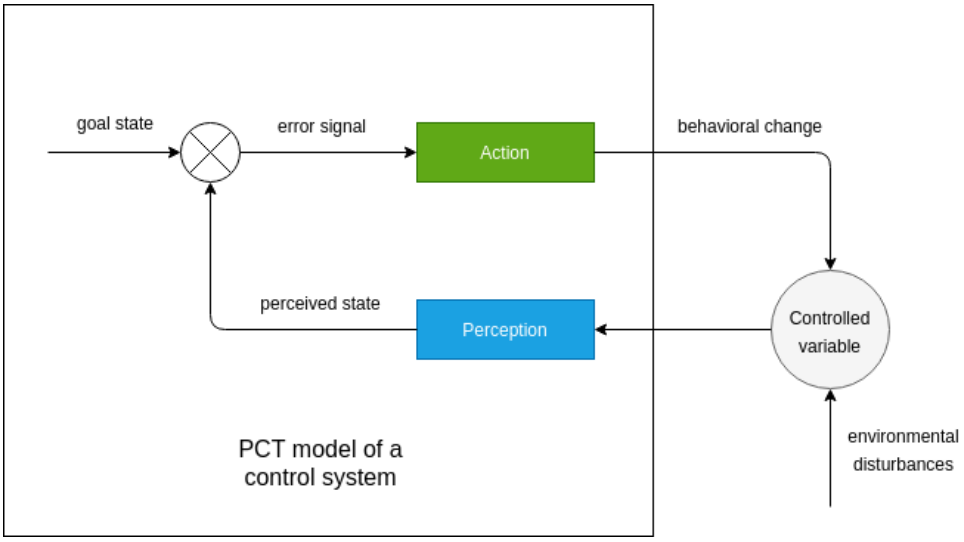

Figure 3: A simple control system with one controller according to PCT (adapted from [20]).

conscious awareness.

A simple PCT model consisting of a single controller is shown in Figure 3. The model involves a closed control loop where an error signal is first calculated based on the difference between the internally predetermined goal state and the perceived state that is the current value of the controlled variable. This error signal is then transformed into a behavioral change that, together with environmental disturbances outside the system's control, affects the value of the controlled variable and results in a new error calculation. In other words, the system is directing its behavior based on feedback from perception, continuously attempting to bring the error closer to zero and perceptions closer to the goal. The system may exhibit behavior even if the error is zero in order to maintain it at zero, and the system can appear not to display behavior if non-behavior reduces the error.

According to PCT, cognition involves a hierarchy of controllers [20, 44]. A single PCT controller in such a hierarchy is shown in Figure 4. Each controller receives a combined reference signal as input from a set controllers above and outputs a reference signal to a controller below. An exception to this is the lowest level where direct sensory signals are received as input and motor actions are produced as output. Controllers at different levels in the

hierarchy control different types of perceptual variables, and even abstract concepts can be constructed as combinations of low-level perceptions. Each controller has a memory switch and a perceptual switch that, depending on their settings, enable the controller to be in one of four modes. If both switches are vertically aligned, the unit is passing perceptions upwards in the hierarchy and taking action normally (*conventional control mode*). If the perceptual switch is vertically aligned and the memory switch is not, the unit can pass perceptions upwards without taking action (*passive observation mode*). If it is the other way around, the unit can take action automatically without passing perceptions upwards (*automatic mode*). If neither of the switches is vertically aligned, perceptions retrieved from memory go directly upwards in the hierarchy (*imagination mode*).

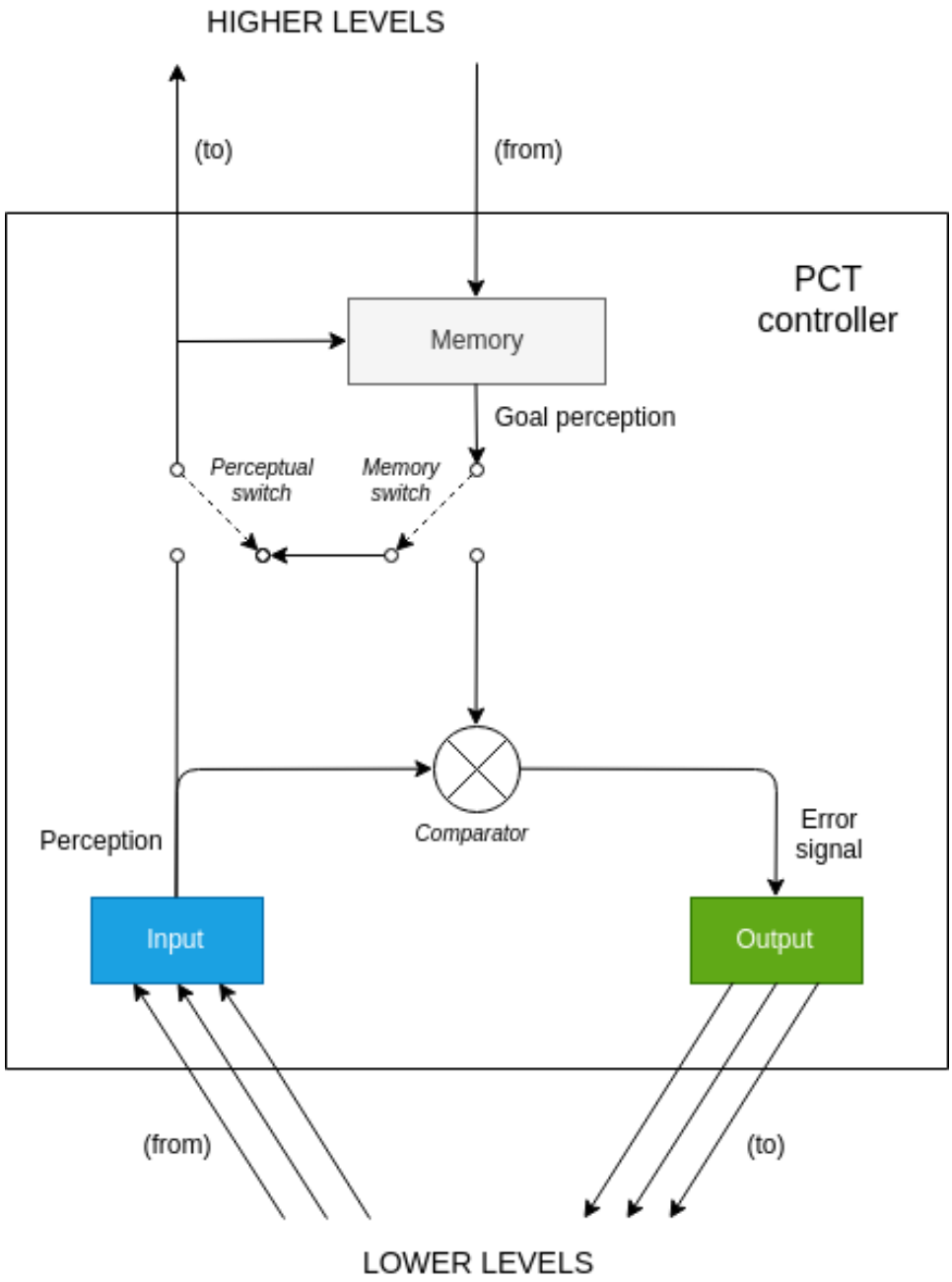

Figure 4: A single controller in a controller hierarchy according to PCT (adapted from [20]).

3 Methods

The amount of existing cognitive architectures is huge yet progress towards achieving a truly general, human-like cognitive architecture seems slow [16]. We found it important to evaluate possible starting points for such an architecture. Therefore, as our research method, we selected a comparative analysis of the standard model and the PCT model based on literature.

Because of the limits of this thesis, we decided to focus on two models only. We chose the standard model and the PCT model because they both strive to be general models of cognition yet differ from each other on a fundamental level, making for an interesting and potentially fruitful comparison. The standard model originates from cognitive science and mostly represents the traditional view of cognition as logical, symbolic computing in sense-think-act cycles [14]. On the other hand, the much older PCT model has its roots in cybernetics and control systems engineering and says cognition is about perceptual control, claiming that the minds of living organisms such as humans resemble analog computers more than digital computers [20]. In addition, the standard model represents a current consensus on cognition that has been synthesized from the views of different disciplines while the PCT model is not widely adopted or well-known. We were interested in finding out how the PCT model challenges the consensus and whether it could offer something new or helpful to the table.

The comparison criteria were gathered from existing literature on requirements for human-like cognitive architectures as well as recent surveys that have identified missing functions in existing cognitive architectures and general AI systems [30, 40, 33, 16, 13, 25]. Because our purpose was to find out how promising the models are as high-level layouts for comprehensive, human-like cognitive architectures, we attempted to select the comparison criteria so that they captured the functions of human cognition as comprehensively as possible. We ended up selecting 10 comparison criteria: *embodiment and perception*, *attention*, *memory*, *learning*, *motivation*, *emotion*, *development*, *language*, *imagination*, and *consciousness*. These criteria are summarized in Table 2. We analysed how both the standard model and the PCT model met each of the criterion. The analysis was performed on a theoretical level; any particular technical implementations of the models such as specific cognitive

architectures were not considered. The analysis was therefore limited to capabilities and restrictions determined by high-level design.

Table 2: The functional comparison criteria used to compare the standard model and the PCT model.

Function	**Description**
Embodiment and perception	Capability to process both internal and external sensory input and produce motor output
Attention	Capability to filter out unimportant sensory input
Memory	Capability to store information as well as modify and retrieve stored information
Learning	Capability to acquire new knowledge that affects behavior in the future
Development	Capability to undergo lasting changes in cognitive structures and processes over time
Language	Capability to learn symbol systems such as language for communication and thought
Motivation	Capability to form goals for behavior based on internal needs and desires
Emotion	Capability to experience and be affected by (simulated) physiological feelings
Imagination	Capability to internally simulate and manipulate sensory information without direct sensory input
Consciousness	Capability to mentally or verbally report own experiences and monitor self

4 Results

In this section, we present the comparative analysis of the standard model and the PCT model that was performed based on the functional criteria defined in the previous chapter: embodiment and perception, attention, memory, learning, development, language, motivation, emotion, imagination, and consciousness. We proceed through the functional criteria one by one. On each function, we explain what it means and how it was addressed by each model. A summary of the results is presented in Table 3.

4.1 Embodiment and perception

As discussed in Background, the embodied approach to cognition emphasizes that the body outside the brain greatly affects cognitive processing in humans. A human-like model of cognition should therefore include sensorimotor capabilities for *perception* and *action* [40]. Perception means the gathering of sensory information from the external environment, and action means the execution of motor actions in the environment. Having better capabilities to sense and act makes it possible to build more complex understanding of the world with more potential for development. However, it should be noted that humans with sensorimotor disabilities do not have significant impairments in higher-level cognitive processing [33]. What seems to be important is the ability to interact with the environment in some manner, not how the body precisely functions and looks like. In order for a model to be fully embodied and human-like, it must also simulate some kind of inner physiological state and the capabilities to perceive and regulate it in addition to the external environment. In living organisms, this internal perception is called *interoception* and the regulative physiological process it is needed for is called *allostasis*. By allostasis, the body tries to achieve *homeostasis* - a balanced physiological state.

The standard model includes an input component called *Perception* that processes sensory information arriving to the system from the external world [14]. This component can consist of several sensory modalities such as vision, audition and touch, each taking in certain type of knowledge, translating it into conceptual form via classification or other pattern-recognition mechanisms,

Table 3: A summary of the functions of the standard model and the PCT model.

Function	**Standard model**	**PCT model**
Embodiment and perception	Specialized components for processing sensory input and producing motor output	Hierarchical distributed processing of sensory input and producing of motor output
Attention	An attentional process between working memory and perception	?
Memory	Separate working, declarative and procedural memory components that process symbolic knowledge and associated numerical metadata	Hierarchical distributed associative memory that stores perceptions and alters reference signals
Learning	Several mechanisms for learning both symbolic knowledge in memories and sub-symbolic knowledge in perception and motor components	Associative learning of perceptions and reorganization that modifies the hierarchy
Development	?	Hierarchy modifies itself over time from initial minimal configuration
Language	Assumed to emerge over several cognitive cycles from the primitives of the model	Words and grammatical rules processed as sequence and program perceptions, meaning observable by higher levels through triggered memory associations
Motivation	?	Reference signals motivate lower levels, initial intrinsic physiological and abstract needs on which higher-level goals build
Emotion	?	Somatic branch in the hierarchy converts errors to physiological changes that are monitored by the reorganizing system
Imagination	?	Imagination mode enables manipulation of perceptions in memory at any level without triggering action
Consciousness	?	The reorganizing system can focus on different parts of the behavioral hierarchy for reorganization, monitoring and testing

and finally passing it to its specific buffer in working memory. The model also includes an output component called *Motor* that converts symbolic knowledge in working memory into suitable actions in the external world. This component can also consist of several modalities for different body parts such as legs and head, each taking in knowledge placed in its own working memory buffer and translating it into specific actions. However, it is unclear how or if internal physiological processes and interoceptive capabilities are taken into account in the standard model.

In the PCT model, perception and action are central concepts [20]. Unlike in the standard model, there are several components (controllers) that process input and output signals. Only the lowest layer in the hierarchy is in direct contact with the body, and each higher-level controller 1) receives perceptual signals coming from the layer below and passes them upwards; 2) receives a reference signal from the layer above, calculates the error signal between it and its current perceptual signal, and passes the error signal downwards as a reference signal. Controllers at different levels in the hierarchy control different types of perceptions, and all abstract high-level concepts are constructed as combinations of low-level perceptions. Currently, nine different types of perceptions have been proposed, ordered here from the lowest to highest: intensity, sensation, configuration, transition, sequence, relationship, program, principle, and system concept. At the lowest level, the variables are direct *intensities* such as chemical or mechanical effects that compose *sensations* such as taste and pressure. Objects correspond to *configurations* that consist of multiple sensations. *Transitions* are perceptions of movements of lower-level perceptions, *sequences* are lower-order perceptions in certain order (i.e. events), and *relationships* capture relations between events. *Programs* are hierarchical structures resembling computer programs that are composed of sequences of relationships between events as well as choice points. Examples of program perceptions are recipes and tasks. *Principles* are heuristics that guide the selection and execution of programs. *Systems* are even more abstract concepts such as society, government or family that are composed of multiple principles. However, there are missing details in the PCT model related especially to the higher level perceptions. Interoception is enabled by a *somatic* branch in the controller hierarchy that interacts with the behavioral branch. The behavioral branch at its lowest level controls by activating muscles while the somatic branch at its

lowest level controls by changing the physiological state of the system to allow the action to be taken, for example by increasing heart rate in a living system. When an error occurs in some high-level controller, changes cascade downwards in both the behavioral and the somatic branch. When they reach the bottom of the somatic branch, they are converted into physiological changes that are then perceived by a separate *reorganizing control system*. The reorganizing control system is explained in the section on learning.

4.2 Attention

Attention is a cognitive function that filters out non-relevant perceptual information coming through different sensory modalities [40, 27]. Attention is believed to involve three sets of processes: the first one sustains the alertness or arousal required for attention, the second one orients attention to a specific sensory modality or location, and the third one performs actual target detection and functions as the gateway through which information reaches awareness. In a developmental and human-like model, there should be certain attentional biases such as paying special attention to objects exhibiting biological motion (people) and to objects or general spatial regions to which action is directed.

In the standard model, there is an attentional mechanism that prevents non-relevant knowledge from passing into working memory [14]. In addition, working memory can pass knowledge such as expectations down to the perception component, affecting the translation of perceptual information in short-term. However, the exact details of these mechanisms are unspecified.

In the PCT model, attention is not specifically addressed [20]. It is unclear if attention is directed by the reorganizing control system, and if so, how. The reorganizing system is discussed in the section on learning.

4.3 Memory

Memory is a cognitive function related to storing and retrieving information [40]. Memory enables learning, helping the system better predict the future and adapt to new situations. Memory can be split into *declarative memory*

and *procedural memory*. Declarative memory stores knowledge related to things or facts while procedural memory contains knowledge about actions or skills. Declarative memory can be further divided into *episodic memory* that stores life experiences (knowledge about what, when, where) and *semantic memory* that stores general knowledge about the world such as concepts and their relations.

The standard model specifies three types of memories [14]. *Working memory* is a temporary workspace where symbols retrieved from long-term memory and the perception component are dynamically combined into larger structures and manipulated according to commands of the procedural memory. These structures may contain partial solutions to problems and short-term knowledge about, for example, the current task, environment and progress in goals. The contents of the working memory at any given time can be seen as representing the "mental" state of the system at that time. Working memory also includes buffers where knowledge retrieval requests to other components and their results are stored. *Procedural long-term memory* stores knowledge about internal (mental) and external (motor) actions typically in the form of condition-action rules. In each rule, the condition side is a symbolic pattern that procedural memory checks against the content of the working memory. If the pattern matches, the action side is executed. The action may modify working memory in some way, including the buffers when a request is initiated to other components. The standard model does not specify how to handle multiple rules with matching patterns, but in any case a single action should be selected with influence from numeric metadata. *Declarative long-term memory* stores factual or semantic knowledge in a graph structure where nodes are concepts and edges their relations. Some additional numerical metadata is involved within each concept, for example measures of how recently it has been accessed and how similar it is to other concepts. This metadata affects how and when the concept is retrieved. Retrieval occurs when a cue is placed in the working memory buffer by procedural long-term memory. Declarative long-term memory can also store the system's past experiences in a conceptualized form, for example a history of the states of working memory. This is called episodic memory. Declarative memory would then be divided into episodic memory containing past experiences and semantic memory containing facts.

In the PCT model, there is an associative memory that stores perceptions and is distributed across the controller hierarchy [20]. In each controller, a reference signal received from above is always augmented by the controller's memory of past perceptions, but the four different settings of the perceptual and memory switches determine whether the augmented (final) reference signal goes directly back up or passes down to be compared with the perceptual signal received from below, affecting behavior. The settings also determine whether the perceptual signal received from below goes upwards and is stored in memory or whether it only goes through the comparator to affect behavior. An unsolved question is where the reference signals of the highest level that controls system perceptions come from.

4.4 Learning

By learning new information, a system can better predict the future and adapt to new situations [40]. Preferably more than one learning mechanism should be included in a cognitive architecture. Humans seem to utilize at least three types of learning: *supervised learning*, *unsupervised learning* and *reinforcement learning*. In supervised learning, the system is provided with correct labels or answers to input data. Learning occurs by figuring out the pattern between the input and output that allows the system to predict correct answers to new, unseen input. In humans, this kind of learning happens via teaching and imitation. In unsupervised learning, the correct answers are not given and the task is to find statistical patterns in the input instead. For example, sensorimotor capabilities can be incrementally and autonomously learned by creating mappings between executed actions and changes in sensory input. Human infants do this in at least two ways: motor babbling and goal babbling. Motor babbling is the execution of seemingly random motor actions and sensing how they affect the environment. Goal babbling means repeatedly trying motor actions in an attempt to reach a certain goal such as reaching or grasping. By building sensorimotor mappings, the infant also starts to understand the separation between their own body and the environment as well as the possible ways to use different objects. In reinforcement learning, a reward follows desired behavior and then encourages that behavior in the future. For example, a child may learn to repeat a certain behavior because it elicits a positive emotion.

Human cognition is assumed to include a process called *representational redescription* by which knowledge representations already present in the mind may transform from implicit and low-level to more explicit and abstract, becoming more accessible, inspectable and explainable to the mind [11]. At a certain moment in time, the mind may contain many different (and possibly redundant) representations of the same type of knowledge, and different types of knowledge may be in different phases of the explicitation process. Some knowledge might stay implicit forever, and not all learned knowledge is initially represented in implicit form. Nevertheless, the capability for explicitation exists in the mind and the process can be triggered by several factors. In this view, the mind not only learns new knowledge but also reuses existing knowledge in new ways. The capability to learn both implicit and explicit knowledge and have interactive processes between them has also been defined as one desired feature of a cognitive architecture [30].

In the standard model, there are learning mechanisms that modify information in the components of the model as the system gathers experiences [14]. Each component except working memory has at least one learning mechanism. Procedural long-term memory has two learning mechanisms, one for composing new rules based on the execution pattern of previous rules and another based on reinforcement learning that modifies the logic of rule selection when multiple rules have matching conditions. Declarative memory also has two mechanisms: one for adding new concepts, relations and metadata and another for modifying existing ones. Sub-symbolic (numeric) information in perception and motor components is also assumed to be learnable. This kind of learning involves the creation and modification of patterns used in the translation process that converts sensory information into symbols and symbols into motor actions. The standard model therefore includes capabilities for learning both implicit knowledge "closer" to the senses and explicit symbolic knowledge, and interaction between them occurs through the translation function, which could be implemented as a neural network, for example. However, exact details of all learning mechanisms such as their type (supervised, unsupervised or reinforcement learning) or techniques to implement them are not specified in the standard model.

In the PCT model, the most essential learning mechanism is *reorganization* by which the controllers and their properties change either randomly (*trial-*

and-error) or by "conscious" effort [20, 23]. The reorganizing system can be thought of as a separate control system that is driven by *intrinsic error*, attempting to keep certain physiological *intrinsic perceptions* such as body temperature and nutritional state at inherited reference values. Instead of outputting behavioral changes, the reorganizing system outputs changes in the controllers and parameters of the behavioral control hierarchy. These changes could include adding or removing controllers at new or existing levels, modifying the reference signal of a controller, changing the connections between controllers situated at different levels, and so on. Reorganization can be triggered when there is a persistent *conflict* between two controllers that give conflicting reference signals to the same lower-level controller. The lower-level controller is still able to meet its goal because the effects of the incoming reference signals are summed together, resulting in a single reference signal. However, both of the higher-level controllers find that their own goal is not met; they form a sort of "deadlock" and essentially become useless. Although reorganization resembles traditional reinforcement learning, there are differences. In reinforcement learning, the probability of a certain behavior increases the more it becomes associated with a positive consequence, i.e. reward. In essence, it is implied that the reward in some way causes (reinforces) behavior. PCT on the other hand says that a reward only stops or slows down a reorganization process that indirectly causes behavioral changes. Reorganization also explains the phenomenon of *transfer* in human learning where knowledge gained in one task is used to solve another unrelated task. In the PCT model, transfer naturally occurs as the system is capable of controlling the same perceptual variable at different reference levels despite varying situations and disturbances. Unlike in the standard model, there is not a translation function or a single level in the hierarchy through which implicit perceptual knowledge is translated into explicit symbolic knowledge; knowledge is more implicit in the initial levels of the hierarchy and goes through an explicitation process over time as higher layers able to perceive e.g. symbolic knowledge are created.

4.5 Development

When discussing learning machines, Turing thought it would be easier to create a child-like AI capable of developing into a mature, adult one instead

of trying to understand and mimic complex adult AI [36]. In Turing's vision, the child AI would initially resemble a blank notebook equipped with very minimal mechanisms that enable further development, and over time, the notebook would be filled as the child AI is educated through rewards and punishments. Turing said this would require one to: "...provide the machine with the best sense organs that money can buy, and then teach it to understand and speak English. This process could follow the normal teaching of a child. Things would be pointed out and named, etc.".

A key feature of human cognition indeed is that it involves developmental processes in addition to learning processes, but this has rarely been taken into account in the design of cognitive architectures [40]. While learning occurs as the agent interacts with the environment and acquires new knowledge, development is the result of the system interacting with itself and changing its structure and processes [38].

The standard model does not include capabilities for fundamentally changing the components and processes of the model over time [14].

In the PCT model, the full control hierarchy with 11 levels of controllers is not supposed to be present initially in the beginning of the system's lifetime [20]. However, there may be some built-in structure or blueprints for these levels so that they do not need to be created from scratch. Development is then only about filling pre-existing perceptual categories with specific examples. Initially, the model includes the reorganizing control system and a simple behavioral control system that consists of reflexive "preprogrammed" behaviors such as shivering and suckling. At the beginning, there is also constant unconscious trial-and-error reorganization in the hierarchy that can manifest as motor babbling, for example. Reorganization stops after the reorganizing system perceives that the intrinsic error has been reduced to zero by some behavioral pattern, making that pattern persist until the intrinsic state for some reason becomes unstable again. The pattern may still be poor in many ways, causing errors in the behavioral control hierarchy, for example, but the reorganizing system does not care about the nature of the behavior as long as it balances the intrinsic state. Eventually, the reorganizing system also starts to use conscious efforts to direct where reorganization takes place.

4.6 Language

In the field of cognitive-functional linguistics, language is not seen as a separate cognitive module but as interwoven with all cognitive systems, and the functions of language are emphasized instead of merely its form [34]. Language has two functions: *semiological* (thought) and *interactive* (communication). Young children first communicate with gestures such as pointing and eventually acquire language which also becomes a tool for thought. Each language is a collection of symbolic *structures*, each structure a combination of different kinds of symbolic *elements*: words, markers on words such as plurals, order of words and intonation. Some general structures may contain only slots for words, and abstract structures may contain only word categories such as verbs, nouns and pronouns in certain order. A structure does not contain meaning; instead, it evokes the construction of meaning, a process of *conceptualization* that involves several sources of information such as knowledge of grammar stored in memory and the current environmental or psychological context. Language is used to conceptualize many kinds of mental experiences such as sensory, emotional and motor experiences in addition to abstract ideas.

In the standard model, it is assumed that language processing capabilities among other more complex capabilities emerge over several cognitive cycles from the model's existing structures and processes without a need for specific ones for language [14]. However, the model allows for primitives specific to language processing to be added, for example a structure in working memory that processes auditory verbal information also known as the phonological loop.

In the PCT model, words or symbols are assumed to be no different from other perceptions [20]. There is no specific level for symbolic perceptions. For example, words belong to sequence perceptions controlled by the fourth level. By the associative nature of memory in controllers, words - as any other perceptions - trigger associated lower-level non-word perceptions such as configurations (objects), events and relationships which then travel upwards in the hierarchy and become observable as the meaning of the words. Grammatical rules, on the other hand, are assumed to belong to program perceptions controlled by the seventh level. Program perceptions are hi-

erarchical structures resembling computer programs that are composed of lower-level perceptions such as *relationships* between *sequences*. Programs also contain decision points or IF statements where the current perceptual state is compared against the desired goal state. If the goal has been achieved, the program exits; otherwise, it continues. In a program perception, the relationships are not between actual behavioral sequences (events) but *word* events, i.e. word perceptions, which are transformed into behavior.

4.7 Motivation

Motivation is the underlying force that drives actions and development [40]. In addition to intrinsic physiological needs such as nutrition, water and sleep that are important for maintaining homeostasis, an *exploratory motive* - in other words, curiosity - is thought to be innate in humans and present at birth. It manifests as a tendency to search for and pay attention to novel stimuli and test limits of actions. Like scientists, infants are curious to explore the world by trying new ways of doing things and learning from the results. This exploration is driven by a need to maximize progress in learning, meaning experiences that most reduce uncertainty in the predictions of action consequences are chosen, and this reduction of prediction error is intrinsically rewarding - not only as a means to achieving a certain goal [26]. Experiences that are too boring or difficult, i.e. experiences that are already easy to predict or have a slow learning rate are not as interesting to infants as those with a fast learning rate. What is notable is that the stage-like developmental trajectory typical to humans can be simulated with this simple exploratory learning mechanism without needing to take into account dependencies to e.g. physiological maturation of the brain. Similarities and differences in developmental trajectories can also be explained; similarities are due to the same general motivational process and differences are due to variability in the environment and experiences offered by it. These findings support the neuroconstructivist view that the human mind is preprogrammed only with certain general biases and mechanisms [11]. The exploratory motive is sometimes separated from a *social motive* to pay attention to, interact with, and imitate other people. However, it is not difficult to see how motivation for learning social skills could also arise from the simple exploratory motive or physiological needs.

In the standard model, there currently is no specific mention of motivation [14]. Working memory includes knowledge about current goals among other content, and an action whose condition matches the content is selected and executed. However, it is unclear what these goals are based on, how completely new goals could be created, and how or if any intrinsic, exploratory or social motives could be simulated.

In the PCT model, reference signals passed downwards by the higher levels in the hierarchy work as *motivators* for the lower levels [20]. Simply, a system implementing the model is motivated to behave in a certain manner because they are controlling some internal perceptual variable so as to keep its value in line with a reference value. In the initial hierarchy that is present in the beginning of the system's lifetime, there are inherited or "pre-coded" controllers that control intrinsic perceptual variables (also called *essential variables*) related to physiological needs by random or reflexive behavior. In addition, there are assumed to be certain intrinsic abstract needs or motives such as curiosity and a "drive to competence", a natural drive by part of the reorganizing system to improve the quality of the whole control system by attempting to minimize its total error. All higher-level goals build on top of these low-level inherited goals as the hierarchy grows.

4.8 Emotion

Several theories and definitions of emotion exist. According to the theory of constructed emotion (also known as conceptual act theory), emotions are psychological constructions or learned conceptualizations of physiological sensations [2]. For example, an emotion word such as fear refers to a conceptual category and not to any universal, basic emotion that is hard-wired in the brain. As discussed in the previous section about language, a conceptualization or "meaning-making" of events involves several sources of information such as incoming sensory information, previous conceptualizations (memory) and the physiological state. When previous conceptualized events of fear are used to conceptualize current events, fear is experienced. The theory of constructed emotion challenges classical emotion theories by arguing that emotions are not reactions to the internal and external environment as much as they are constructions of it. The reason why similar emotion concepts

exist in different cultures is because they happen to serve similar functions in those cultures. It has been noted that the younger a child is, the more their choices and behavior are driven by emotion [1].

In the standard model, there is currently no mechanism for emotions [14].

In the PCT model, emotions result from the physiological changes caused by the somatic control branch which was explained in the section on perception [20]. The reorganizing system that monitors the physiological state may not be aware of the root cause of the emotion, i.e. the location of the error in the hierarchy, and it may seem like the emotion appeared out of nowhere. The larger the error, the more negative the emotion. It should be noted here that the size of the error does not necessarily correspond to the amount of change in the controlled variable. This is because of a controller-specific parameter called *loop gain* that determines how much the incoming perceptual signal is amplified. In other words, this mechanism gives different importance to different perceptions (and errors). Changes in perceptions that do not matter do not result in large errors and strong negative emotions.

4.9 Imagination

It has been pointed out that a crucial ability that enables better control of the environment and oneself, therefore improving adaptability, is the ability to predict what could happen under different conditions, in other words *imagination* [42]. Imagination is a process by which information that is not directly received through the senses (i.e. is stored in memory) is used in mental simulation of possible events and actions [21, 40]. It is suspected that in human infants, mental simulation initially occurs in dreams and gradually becomes reliable enough to be used as a tool in waking life. Imagination allows the system to predict consequences of different actions without needing to pay the physical cost of action and experience those consequences in the real world.

In the standard model, imaginative capabilities are not addressed [14].

In the PCT model, there is a mechanism for imagination [20]. A controller is in imagination mode when neither its memory switch nor its perceptual

switch is vertically aligned. In this mode, the incoming reference signal (altered by memory) goes directly back up the hierarchy without passing through the controller's comparator. Theoretically, this enables the ability to internally manipulate perceptions at any level without causing error calculations and behavior. Thought can be seen as a form of imagination where the perceptions imagined are word perceptions that belong to the sequence perceptions controlled by the fourth level. The PCT model also explains dreaming: in dreaming, the control hierarchy is optimized "off-line".

4.10 Consciousness

Consciousness has two, partly overlapping definitions [5]. First, consciousness refers to a system's capability to in some way globally access the information it processes which enables further processing, for example mental and/or verbal reporting. Second, it means the system's capability to process information about itself, also known as *metacognition*.

In the standard model, there is no mention of any mechanism for consciousness [14].

In the PCT model, the capability for consciousness is assumed to come through the reorganizing system [20]. The reorganizing system is able to perceive the behavioral hierarchy itself in a selective manner, to monitor the perceptions processed by certain parts of it and even inject test perceptions into controllers in order to see the effects on behavior. The reorganizing system also directs the reorganization (learning) process this way. The parts of the hierarchy that are not monitored function in unconscious mode. However, there are missing details related to what exactly determines the behavior of the reorganizing system including the directing of consciousness.

5 Discussion

In this section, we summarize the results of the previous section, discuss the benefits and problems of the models, highlight certain problems related to our study, and offer future directions for research.

The standard model in its current state does not seem to address certain cognitive functions such as motivation, emotion, development, imagination and consciousness. The functions addressed by the standard model - perceptual and motor capabilities, attention, memory and learning - are those that have traditionally received more attention in cognitive architecture research. The PCT model seems to address all functions except attention, therefore being more comprehensive on a theoretical level. However, the specifics of the reorganizing system responsible for many of the cognitive functions are unclear. With its hierarchical and distributed structure and parallel processing, the PCT model has some resemblance to connectivist models although it is based on mathematical control theory instead of statistical techniques such as ANNs. The role of the body outside the brain is central in the PCT model, and interoceptive capabilities in addition to exteroceptive capabilities are taken into account unlike in the standard model. It is unclear how a system implementing the standard model could independently form its own goals and act to achieve them. What would motivate the system to, for example, explore and learn social skills and language?

5.1 Development

It is known that in humans, many cognitive capabilities develop over time and build on top of previous knowledge from birth to adulthood, especially during the first few years of life. The importance of a developmental approach to cognitive science has been emphasized [11, 22], and desirable features for developmental cognitive architectures have been proposed [40]. In addition, the need for some kind of a "developmental start-up software" has been recognized in light of the shortcomings of systems based purely on neural networks [15]. The idea in developmental architectures is that their structure, processes and knowledge are initially very minimal like those in the mind of a newborn child. However, they develop over time as the child

AI system is allowed to interact with its environment, gather experiences and be educated by humans in either real life or a virtual world. Gradually, its cognitive architecture matures into an adult-like cognitive architecture. This capability of a cognitive architecture to modify its own structure and processes is called *constitutive autonomy*. [40]. It is related to the requirement for *bio-evolutionary realism* that says a cognitive architecture should not defy the biological and evolutionary history of living organisms including humans.

Some regard it problematic that the standard model assumes certain structures and processes of the mind are innate and fixed while all learned abilities are represented as knowledge and skills, i.e. as content of a fixed architecture [32]. In reality, the distinction between innate and learned cognitive abilities is not so clear, and most of what we think is innate may actually be learned and changeable. The standard model does not have developmental capabilities; it mainly seems to model the end result of development, i.e. the fixed structure and processes of an adult-like mind with mechanisms for learning new knowledge.

In the PCT model, the behavioral hierarchy starts off very minimal but scales over time as new controllers and completely new layers are created in a process that seems at least functionally equivalent to the development process of human cognition. This makes the PCT model perhaps more suitable as a developmental and thus human-like model. In fact, there already is a popular book on baby development inspired by PCT [37]. Most of the research related to developmental, human-like AI has been carried out in the field of developmental robotics where the focus has been on low-level sensory-motor capabilities instead of reasoning, planning and other higher-level cognitive capabilities [40]. These robotic systems are often built by integrating off-the-shelf modules in a pragmatic manner without any general cognitive architecture as a basis. We think the PCT model could have potential to provide such a general architecture for developmental robots.

5.2 Cognitive realism

Another requirement for a cognitive architecture is *cognitive realism* meaning its features should not be in conflict with what we know about human

cognition [30]. However, it does not necessarily need to accommodate every slight difference that may exist between humans at least what comes to structure and procedures. An architecture should be grounded in theories of the basic building blocks of human cognition with individual differences naturally emerging from the variable content of the architecture that varies between different instantiations.

Both the standard model and the PCT model have been inspired by theory instead of practice, the former because it was intended as a theory about living organisms first and foremost and as a platform for AI applications second. Although the PCT model seems more comprehensive and is empirically supported as well [20], some people may doubt whether its nontraditional view of cognition as control of perceptions is realistic. After all, the model is not currently widespread and accepted in psychology or cognitive science and does not represent any kind of consensus unlike the standard model. It also has some crucial missing details. For example, it is not clear how certain aspects of it such as the reorganizing system, the higher levels of the control hierarchy and the somatic control branch could be implemented in an actual cognitive architecture. It may turn out to be challenging or even impossible to fully implement the PCT model with currently known techniques and see whether its assumptions hold in practice although initial steps have been taken. For example, there have been efforts towards building a PCT-inspired robotics architecture [44] and a computational framework that supports interpretation of perceptions and communication [23]. The standard model, on the other hand, has been synthesized from existing cognitive architectures and thus is founded on cognitive theories. Despite its theoretical deficiencies, the standard model and the traditional sense-think-act view may be more acceptable by the research community and translate more easily into practical software applications. The standard model is understandably incomplete because it represents a consensus, and reaching consensus on the missing cognitive functions could take time. Because it is not complete or even necessarily intended to be, it also leaves more room for the requirements of specific applications.

5.3 Neural plausibility

It has been regarded as problematic that the standard model represents a horizontal model where a certain level of abstraction (the cognitive level) is deemed as the starting point for modeling cognition and the levels below (the neural levels) are ignored simply as implementation [32]. This view that the mind can exist without the brain like a software independent of the hardware it runs on is called *computational functionalism*, and arguments for and against it have been presented. Developing multi-level models that take into account all levels of abstraction instead of horizontal models has been recommended.

Unlike the standard model, the PCT model is intended to be functionally equivalent to the nervous system, meaning it is in accordance with the high-level functions of the neural level although does not share a similar implementation as the brain [20]. The PCT model seems to be multi-level: in addition to being neurally plausible, it is capable of explaining functions at the levels of deliberate action. As a singular basis for a cognitive architecture at the levels of deliberate action - rational, cognitive and social - connectionism seems insufficient [6, 15]. For example, it is not clear how a purely connectionist architecture could account for the compositional and causal nature of thought.

5.4 Autonomy and PCT

Autonomy is one dilemma identified in the design of cognitive architectures [39]. At one end, the architecture can be allowed to be fully autonomous so that its behavior cannot be controlled from the outside. At the other end, the architecture can have zero autonomy and be completely dependent on the commands of other agents. It would naturally seem important that in a human-like cognitive architecture, human-like autonomy and self-determination emerges. However, there may be situations where programming the PCT system by hand is necessary, for example when it is being practically evaluated and natural development would take too much time. The problem with developmental models such as the PCT model is that the environmental interaction required for development - and therefore

development itself - can not necessarily be simulated in a shorter time than is naturally possible for humans [40]. Of course, different practical applications require different capabilities. For example, capabilities that are on the level of a typical one-year-old could be enough for some practical applications such as companion robots. However, it seems highly impractical to teach every instantiation of a cognitive architecture implementing the PCT model for even that long. One solution could be to create some kind of virtual worlds; worlds that resemble the real world as closely as possible but where the interactions can be sped up considerably. This is definitely a challenging task.

The question is how manual programming of the system could be done both efficiently and in a way that would produce the same results as "natural" development by reorganization without unintended side consequences. What if unintended consequences do occur because of natural development or manual changes in a PCT system? For example, when a single high-level reference signal is changed, complex changes may cascade down the entire hierarchy. What if the system starts exhibiting behavior that is harmful to others in order to achieve its desired perceptions? What crosses the line, when is human intervention justified, and could a human understand the system well enough to control it? PCT emphasizes that other people cannot be predictably controlled unless one knows all the perceptual variables they are controlling and their relationships, i.e. the workings of their whole control hierarchy. In the same way, if one were to predictably affect changes in an artificial control system, a complete understanding of its workings would be required. Although the basic principles of the PCT model may sound simple and a very restricted part of the control hierarchy may be understandable to a human, the hierarchy could quickly scale to such proportions that it would be difficult to comprehend it. What kind of ethical problems are related to creating AI systems that are human-like yet controllable? If PCT is correct and humans are nothing but organic control systems, would there be any difference between artificial control systems and us significant enough to justify different treatment?

5.5 Predictive processing and PCT

Neuroscience has traditionally offered insights on only a limited set of cognitive functions such as sensorimotor capabilities but recent promising theories such as the *predictive processing* (PP) theory are argued to be capable of explaining many others as well [16, 3]. In cognitive neuroscience, the PP framework has gained a lot of traction as a possible unifying theory of perception, action and cognition [3, 9, 28]. PP suggests the brain is essentially a probabilistic prediction machine that follows Bayesian principles. The brain's purpose is *prediction error minimization*; knowledge about the world is encoded in the brain as probability density functions, and the brain constantly generates hypotheses (predictions) about oncoming perceptions and attempts to minimize errors between what it predicts and what it actually perceives. To do this, the brain uses a generative statistical model that consists of a hierarchy of levels, each level predicting the output of the level below. Predictions arise from prior probabilities of the top level and propagate down to the lower levels that calculate the errors between them and actual inputs. The prediction that best explains reality, i.e. the one with the highest posterior probability and the smallest error, is then perceived. The errors propagate back to the top where they trigger changes in the prior probabilities and cause the rapid generation of new predictions that cancel out the previous ones. The errors also cause structural long-term changes in the generative model itself, i.e. learning. Action and attention are tightly coupled with perception. Action enables *active inference* - the selection of perceptual data that conforms to our internal generative model as an alternative to changing the model. Attention is a mechanism for controlling the precision of errors and therefore their "weight" or influence on the model. Different views exist on the exact mechanisms of prediction error minimization, the most known being *predictive coding* that was originally developed as an efficient strategy for signal processing. According to predictive coding, only the prediction errors propagate back to the top levels and the sensory inputs themselves are ignored. This kind of predictive processing approach that involves predictive coding is called *hierarchical predictive processing*.

Predictive processing may offer an account of perception and action but it is unclear how the hierarchical generative model could produce higher cognitive capabilities such as reasoning and thought that are not tied to perceptions

[41]. Thought is based on combining concepts that represent the world at different timescales and spatial distances as well as abstract concepts such as money, and the question is how or if this kind of capability is possible with the hierarchical generative model of the PP framework. Another problem is related to motivation [12]. However, evidence exists for predictive processing capability in infants and it has even been argued to be the most fundamental learning mechanism in the brain, although its exact workings can change over development [35]. The PP framework and the PCT model have some outward similarities. In PP, the central concept is *prediction* instead of *perception*, and the foundation is Bayesian statistics instead of control theory. In the future, it would be interesting to compare the PCT model and the PP framework in more detail and find out if there is a possibility to synthesize the two.

5.6 On sensory disabilities

Studying humans with sensory disabilities could tell more about what human-like cognition truly means [17]. It is possible that the fundamental structure and processes underlying cognition are simple and general with complexity arising from interaction with the environment. All functionality under a particular cognitive system may not be necessary for a cognitive architecture to exhibit human-level intelligence and be capable of the everyday activities of humans. For example, instead of tackling several sensory modalities or even single complex modalities such as sight and hearing, it could be enough to implement a very minimal set of perceptual capabilities; capabilities that enable some form of two-way communication (teaching and learning) and world-building. This is also in line with the thoughts of Alan Turing:

"We need not be too concerned about the legs, eyes, etc. The example of Miss Helen Keller shows that education can take place provided that communication in both directions between teacher and pupil can take place by some means or other." [36]

5.7 Practical evaluation

In this thesis, we analyzed only a certain set of cognitive functions. For example, we could have separately studied decision-making, problem-solving, social skills, personality, and creativity. In addition, although it may be useful and interesting to compare models that are not similar to each other, in this case there were perhaps too many fundamental differences between the models in terms of theoretical assumptions and level of abstraction that made a reasonable comparison difficult, leaving the comparison somewhat lacking. Furthermore, we were of course not able to uncover the benefits and challenges of the models in practice since the analysis focused only on theory.

Analyzing and comparing how implementations of the models fare in practice would tell more about the true benefits and problems of each model. For example, it would be important to find out whether the implementations are *ecologically realistic*, i.e. able to perform activities that are natural to humans in their everyday life, not only puzzles and games [30]. These activities involve behavior that is fast and reactive, sequential, routine-like, and mostly learned via trial-and-error. In practical evaluation, the way the models scale over time would also become apparent. It would be interesting to find out how large the controller hierarchy of the PCT model could get in real scenarios and how the reorganization process could be implemented efficiently. In the standard model, the translation function between working memory and sensorimotor components might become one possible bottleneck.

6 Conclusion

Contemporary analytic AI research has focused on specialized aspects of cognition such as vision or language processing and deep learning-based techniques for implementing them. However, these kind of AI systems usually have problems with learning abstract, causal and conceptual knowledge, are competent in only limited domains, and can be difficult for humans to understand, communicate with, and trust. A need for more synthetic research that studies the development of more general AI systems that are capable of human-level intelligence has been recognized. In this endeavor, taking inspiration from theories on human cognition may be beneficial. These kind of unified theories of the structure and processes of the mind are called cognitive architectures. However, even though hundreds of these architectures have been developed over several decades, we seem to be nowhere near achieving general AI.

In this thesis, we presented a comparative theoretical analysis of two potential starting points for a comprehensive, general and truly human-like cognitive architecture: the standard model of the mind and the PCT model. The standard model of the mind is a recently proposed layout for cognitive architectures that represents current consensus on human-like cognition across several disciplines. Although statistical learning of mostly sensorimotor capabilities is included in the model, its foundations are in the classical paradigm of cognitive science that models the mind as somewhat computer-like processor of symbols whose output is a function of input. The older and less-known PCT model is based on mathematical control theory and suggests the human mind is composed of a hierarchy of controller units that attempt to control input perceptions by behavioral output, arguing the traditional views on cognition are not applicable to living organisms and therefore human cognition.

The analysis was performed based on functional criteria gathered from literature on human-like cognition, cognitive architectures, and general artificial intelligence. These criteria were embodiment and perception, attention, memory, learning, development, language, motivation, emotion, imagination and consciousness. Our results indicate that while the PCT model seems more comprehensive on a theoretical level and may be especially suitable as a

developmental model, for example as a general architecture for developmental robots, it has some problems that currently make its full implementation difficult. The standard model, on the other hand, may translate more easily into practical implementations.

In the future, additional research is required in order to fill the theoretical gaps of the standard model and the implementational gaps of the PCT model. Their implementations should be practically evaluated with methods that are suitable for general AI systems. This way, their true benefits and challenges would become more clear. In addition, it could be interesting to compare how the PCT model differs from the recent predictive processing framework. Studying cognitive and sensory disabilities could clarify the definition of human-like cognition and what is really required from human-like AI, and ethical problems related to autonomy of human-like AI should be considered.

Outside the traditional, widespread beliefs on what cognition is and what achieving general AI most likely requires, there can be promising, alternative paths. What if the knowledge that not only helps us build general, human-like AI but also understand ourselves can eventually be found in engineering instead of cognitive science?

References

[1] Arsalidou, M. and Pascual-Leone, J.: *Constructivist developmental theory is needed in developmental neuroscience.* npj Scientific of Learning, 1:16016, December 2016.

[2] Barrett, L.F., Wilson-Mendenhall, C.D., and Barsalou, L.W.: *The conceptual act theory: a road map.* In Barrett, L. Feldman and Russell, J.A. (editors): *The Psychological Construction of Emotion*, pages 83–110. Guilford Press, New York, 2014. `http://eprints.gla.ac.uk/112258/`.

[3] Clark, Andy: *Whatever next? predictive brains, situated agents, and the future of cognitive science.* Behavioral and Brain Sciences, 36(3):181–204, 2013.

[4] Dawson, Michael: *Mind, Body, World: Foundations of Cognitive Science.* August 2013, ISBN 9781927356173.

[5] Dehaene, Stanislas, Lau, Hakwan, and Kouider, Sid: *What is consciousness, and could machines have it?* Science, 358(6362):486–492, 2017, ISSN 0036-8075. `https://science.sciencemag.org/content/358/6362/486`.

[6] Fodor, Jerry A. and Pylyshyn, Zenon W.: *Connectionism and cognitive architecture: A critical analysis.* Cognition, 28(1):3 – 71, 1988, ISSN 0010-0277. `http://www.sciencedirect.com/science/article/pii/0010027788900315`.

[7] Grigsby, Scott: *Artificial Intelligence for Advanced Human-Machine Symbiosis*, pages 255–266. June 2018, ISBN 978-3-319-91469-5.

[8] Hern, Alex: *Shotgun shell: Google's ai thinks this turtle is a rifle.* Nov 2017. `https://www.theguardian.com/technology/2017/nov/03/googles-ai-turtle-rifle-mit-research-artificial-intelligence`.

[9] Hohwy, Jakob: *The predictive mind.* Oxford University Press, 2013.

[10] Huitt, William: *The information processing approach to cognition.* Educational psychology interactive, 3(2):53, 2003.

[11] Karmiloff-Smith, Annette: *Précis of beyond modularity: A developmental perspective on cognitive science.* Behavioral and Brain Sciences, 17:693–745, January 1997.

[12] Klein, Colin: *What do predictive coders want?* Synthese, 195(6):2541–2557, Jun 2018, ISSN 1573-0964. https://doi.org/10.1007/s11229-016-1250-6.

[13] Kotseruba, Iuliia, Gonzalez, Oscar J. Avella, and Tsotsos, John K.: *A review of 40 years of cognitive architecture research: Focus on perception, attention, learning and applications.* CoRR, abs/1610.08602, 2016. http://arxiv.org/abs/1610.08602.

[14] Laird, John, Lebiere, Christian, and Rosenbloom, Paul: *A standard model of the mind: Toward a common computational framework across artificial intelligence, cognitive science, neuroscience, and robotics.* 38:13, December 2017.

[15] Lake, Brenden M., Ullman, Tomer D., Tenenbaum, Joshua B., and Gershman, Samuel J.: *Building machines that learn and think like people.* Behavioral and Brain Sciences, 40, Nov 2016, ISSN 1469-1825. http://dx.doi.org/10.1017/S0140525X16001837.

[16] Langley, Pat: *Progress and challenges in research on cognitive architectures.* 2017. https://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/15042.

[17] Leiber, Justin: *Helen keller as cognitive scientist.* Philosophical Psychology, 9(4):419–440, 1996. https://doi.org/10.1080/09515089608573193.

[18] Lieto, Antonio, Bhatt, Mehul, Oltramari, Alessandro, and Vernon, David: *The role of cognitive architectures in general artificial intelligence.* Cognitive Systems Research, 48:1 – 3, 2018, ISSN 1389-0417. http://www.sciencedirect.com/science/article/pii/S138904171730222X, Cognitive Architectures for Artificial Minds.

[19] Marcus, Gary: *Deep learning: A critical appraisal.* CoRR, abs/1801.00631, 2018. http://arxiv.org/abs/1801.00631.

[20] Marken, Richard and Mansell, Warren: *Perceptual control as a unifying concept in psychology.* Review of General Psychology, 17:190, June 2013.

[21] Marques, Hugo Gravato and Holland, Owen: *Architectures for functional imagination.* Neurocomputing, 72(4):743 – 759, 2009, ISSN 0925-2312. http://www.sciencedirect.com/science/article/pii/S0925231208004645, Brain Inspired Cognitive Systems (BICS 2006) / Interplay Between Natural and Artificial Computation (IWINAC 2007).

[22] MELTZOFF, ANDREW N.: *Towards a developmental cognitive science.* Annals of the New York Academy of Sciences, 608(1):1–37, 1990. https://nyaspubs.onlinelibrary.wiley.com/doi/abs/10.1111/j.1749-6632.1990.tb48889.x.

[23] Moore, Roger K.: *PCT and beyond: Towards a computational framework for 'intelligent' communicative systems.* CoRR, abs/1611.05379, 2016. http://arxiv.org/abs/1611.05379.

[24] Newell, Allen: *Unified theories of cognition.* Harvard University Press, 1994.

[25] Nirenburg, Sergei: *Cognitive systems: Toward human-level functionality.* AI Magazine, 38:5, December 2017.

[26] Oudeyer, Pierre Yves and Smith, Linda B.: *How evolution may work through curiosity-driven developmental process.* Topics in Cognitive Science, 8(2):492–502. https://onlinelibrary.wiley.com/doi/abs/10.1111/tops.12196.

[27] Petersen, Steve and Posner, Michael: *The attention system of the human brain: 20 years after.* Annual review of neuroscience, 35:73–89, April 2012.

[28] Rao, Rajesh and H. Ballard, Dana: *Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects.* Nature neuroscience, 2:79–87, February 1999.

[29] Shevtsov, S., Berekmeri, M., Weyns, D., and Maggio, M.: *Control-theoretical software adaptation: A systematic literature review.* IEEE Transactions on Software Engineering, PP(99):1–1, 2017.

[30] Sun, Ron: *Desiderata for cognitive architectures.* Philosophical Psychology, 17(3):341–373, 2004.

[31] Sun, Ron: *Potential of full human–machine symbiosis through truly intelligent cognitive systems.* AI & SOCIETY, Nov 2017. https://doi.org/10.1007/s00146-017-0775-7.

[32] Taatgen, Niels: *Cognitive architectures: Innate or learned?* pages 476–480, November 2017.

[33] Thill, Serge: *What we need from an embodied cognitive architecture.* In *Cognitive Architectures :*, number 94 in *Intelligent Systems, Control and Automation: Science and Engineering*, pages 43–57. 2019, ISBN 978-3-319-97549-8.

[34] Tomasello, Michael: *The new psychology of language: Cognitive and functional approaches to language structure*, volume 1. Psychology Press, 2014.

[35] Trainor, Laurel J.: *Predictive information processing is a fundamental learning mechanism present in early development: Evidence from infants.* International Journal of Psychophysiology, 83(2):256 – 258, 2012, ISSN 0167-8760. http://www.sciencedirect.com/science/article/pii/S0167876011003874, Predictive information processing in the brain: Principles, neural mechanisms and models.

[36] Turing, Alan M: *Computing machinery and intelligence.* In *Parsing the Turing Test*, pages 23–65. Springer, 2009.

[37] Vanderijt, Hetty and Plooij, Frans X: *The Wonder Weeks: How to Stimulate Your Baby's Mental Development and Help Him Turn His 10 Predictable, Great, Fussy, Phases Into Magical Leaps Forward.* Kiddy World Promotions, 2010.

[38] Vernon, David: *Enaction as a conceptual framework for developmental cognitive robotics.* Paladyn, 1(2):89–98, Jun 2010, ISSN 2081-4836. https://doi.org/10.2478/s13230-010-0016-y.

[39] Vernon, David: *The Architect's Dilemmas*, pages 59–70. Springer International Publishing, Cham, 2019. https://doi.org/10.1007/978-3-319-97550-4_5.

[40] Vernon, David, Hofsten, Claes von, and Fadiga, Luciano: *Desiderata for developmental cognitive architectures.* Biologically Inspired Cognitive Architectures, 18:116 – 127, 2016, ISSN 2212-683X. http://www.sciencedirect.com/science/article/pii/S2212683X16300822.

[41] Williams, Daniel: *Predictive coding and thought.* Synthese, Mar 2018, ISSN 1573-0964. https://doi.org/10.1007/s11229-018-1768-x.

[42] Williams, Daniel: *Predictive minds and small-scale models: Kenneth craik's contribution to cognitive science.* Philosophical Explorations, 21(2):245–263, 2018. https://doi.org/10.1080/13869795.2018.1477982.

[43] Ye, Peijun, Wang, Tao, and Wang, Fei Yue: *A survey of cognitive architectures in the past 20 years.* IEEE transactions on cybernetics, PP, August 2018.

[44] Young, Rupert: *A general architecture for robotics systems: A perception-based approach to artificial life.* Artificial Life, 23(2):236–286, 2017. https://doi.org/10.1162/ARTL_a_00229, PMID: 28513206.