The Architect's Dilemmas

David Vernon
e-mail: vernon@cmu.edu

Abstract The creation of a cognitive architecture presents the architect with many design choices. Some of these choices come in the form of a dilemma, in which the selection of any option over another entails both benefits and opportunity costs. This chapter highlights three dilemmas that confront the architect when deciding how the key issues of fidelity, embodiment, and autonomy should be addressed and reflected in the design. In each case, it discusses the various options, their roots, and the consequences and costs of choosing one option over another. It concludes by considering these three dilemmas in the context of the stance on cognitive adopted by the editors of this book.

1 Introduction

The design of a cognitive architecture is a daunting undertaking, involving many challenges on a scale that is not always apparent when one embarks on the task. The time and effort involved almost always exceed expectation, sometimes leading to a project that spans decades [2, 22, 46, 47, 35, 13]. The task is made all the harder by the fact that the design options derive from underlying principles, in cognitive science and cybernetics, for example, that are not always evident. Worse still, they often involve choices between two apparently competing options, both of which have elements that are attractive. In this chapter, we highlight three such dilemmas — fidelity, embodiment, and autonomy — and we look at the choices in each case, examining the consequences of choosing one option over another. At the end, we reflect on the choices implied by the characterization of cognition that has motivated this book. We begin by examining the role of a cognitive architecture in the design and implementation of a cognitive system.

David Vernon
Carnegie Mellon University Africa in Rwanda

To appear in Cognitive Architectures, I. Ferreira, J. Sequeira, R. Ventura (eds.), Intelligent Systems, Control and Automation: Science and Engineering, Springer (in press).

2 The Role of Cognitive Architecture

Like architecture in the built environment and system architecture in software engineering, a cognitive architecture captures both abstract conceptual form and details of functional operation, focusing on inner cohesion and self-contained completeness [49]. The goal of creating a complete model is significant. It means that all of the mechanisms required for cognition fall under the compass of a cognitive architecture. These include perception, action, control, learning, reasoning, memory, adaptivity, and prospection. This accords cognition much greater breadth than has been the case in the past, when it was viewed by many as a reasoning and planning filling sandwiched between perception and action. Today, cognition, as a process, and a cognitive architecture, as a framework, are seen to embrace all of the elements required for effective action [28]. Thus, a cognitive architecture reflects the specification of a complete cognitive system, its components, and the way these components are dynamically related as a whole. It provides both an abstract model of cognition and the sufficient basis for a software instantiation of that model [25]. Ron Sun captures this succinctly:

> "A cognitive architecture provides a concrete framework for more detailed modelling of cognitive phenomena, through specifying the essential structures, division of modules, relations between modules, and a variety of other aspects" [45].

A cognitive architecture makes explicit the set of assumptions upon which that cognitive model is founded. Depending on the purpose of the modelling exercise, an issue we will mention below and return to in Section 3, these assumptions are derived from several sources: biological or psychological data, philosophical arguments, or hypotheses inspired by work in different disciplines, such as cognitive neuroscience and artificial intelligence.

In essence, then, the role of a cognitive architecture is to provide a complete model of cognition and to do so at at least two levels of abstraction, setting out the overall process by which cognition produces effective action (in whatever guise that may take) and the detailed computational elements by which that process is effected, including formalisms for knowledge representations and the types of memory used to store them, the processes of reasoning, inference, and prediction that act upon that knowledge, and the learning mechanisms that acquire it.

In a sense, a cognitive architecture captures the top two layers of Marr's three-level hierarchy of abstraction, also known as the *Levels of Understanding* framework [27, 26], i.e., the top level computational theory and, below this, the level of representation and algorithm. At the bottom (third) level, there is the implementation or instantiation of this algorithmic and representational framework: the realization of the cognitive architecture as a working cognitive system.

Once it has been created and instantiated, a cognitive architecture plays a second role, providing the means to validate the assumptions and hypotheses on which the computational model is based, refine their representational and algorithmic foundations, and develop their implementation further.

3 The Dilemma of Fidelity

The model of cognition encapsulated in a cognitive architecture may refer either to natural cognitive agents, to artificial cognitive agents, or to both. The term itself has its roots in cognitive science (a branch of human psychology) and is credited to Allen Newell and his colleagues in their work on a *unified theory of cognition* [32, 33], i.e., a theory that covers a complete range of cognitive issues, such as attention, memory, problem solving, decision making, and learning, from a comprehensive set of perspectives, including psychology, neuroscience, and computer science. Allen Newell and John Laird's Soar architecture [23, 37, 20, 22], John Anderson's ACT-R architecture [1, 2], Paul Rosenbloom's Sigma architecture [38], and Ron Sun's CLARION architecture [45, 47] are all candidate unified theories of cognition. Recognizing the importance of generality and completeness mentioned above, recent work is endeavouring to bring various strands together to create a standard model of mind and a consensus on what must be included in a cognitive architecture in order to provide a human-like mind [24].

However, some cognitive architectures, e.g., [12, 15], make no claim about the biological plausibility of the cognitive architecture, although they often draw inspiration from what is known about cognition in natural systems. Instead, they focus on the practical application of cognitive science.

In effect, there are two reasons to design a cognitive architecture: one is to gain a better understanding of cognition in general and the other is to build artificial systems that have capabilities that are commonly found in humans [17]. The motivation for the first is a principled one, the motivation for the second is a practical one. These two motivations are obviously different, but they are not necessarily complementary. There is no guarantee that success in designing a practical cognitive architecture for an application-oriented cognitive robot will shed any light on the more general issues of cognitive science. Similarly, it is not evident that efforts to date to design general cognitive architectures have been tremendously successful for practical applications.

From the principled perspective, a cognitive architecture is an abstract metatheory of cognition that focuses, as we have mentioned, on generality and completeness [24], drawing on many sources in shaping these architectures, often encapsulated in lists of design principles and desirable features referred to as *desiderata* [45, 18, 21, 52]. The second perspective focusses on the practical necessities of the cognitive architecture and designing on the basis of user requirements. Here, the goal is to create an architecture that addresses the needs of an application without being concerned whether or not it is a faithful model of cognition. These two approaches have been dubbed *design by desiderata* and *design by use case* [51].

The dilemma, then, is this: should a cognitive architecture be a general or a specific framework? Should you focus on discovery or invention? Should you favour fidelity or expediency? These are important design questions because a specific instance of a cognitive architecture derived from a general schema will inherit relevant elements embedded in a well-founded framework, but it may also inherit elements that are not strictly necessary for the specific application domain, yielding an architecture that is more complicated than is necessary for that specific application domain. If your focus is on creating a practical cognitive architecture for a specific application, it may not be appropriate to instantiate a design guided by desiderata; arguably, you are better off proceeding in a conventional manner by designing a system architecture that is driven by user requirements, drawing on the available repertoire of AI and cognitive systems algorithms and data-structures. However, the danger here is that the systems perspective that is crucial to cognitive architectures may not be as well-grounded in firm principles as it needs to be. Conversely, if your focus is a unified theory of cognition, then developing use cases and designing a matching system architecture is unlikely to yield insights on the underlying principles of cognition. You may miss some of the key considerations that make natural cognitive systems so flexible and adaptable, and it is unlikely that you will shed much light on the bigger questions of cognitive science.

4 The Dilemma of Embodiment

Embodiment — or, more specifically, embodied cognition — refers to the role that an agent's body plays in the cognitive function of that agent. Possessing a body, however, does not necessarily mean that an agent is embodied, since that body may play no causal role in the agent's cognitive processes.

The cognitive systems community is divided into two schools: those that think an agent's body plays no causal role and those that think it does.[1] Among those that think it does, there are several stances that vary according to the strength of the assertions they make. The dilemma that confronts the architect designing a cognitive architecture is to select which stance to adopt on embodiment. In the following, we will outline the various stances and the implications of adopting one or another in the design of a cognitive architecture.

The essence of cognitivism, a widely-adopted paradigm of cognitive science, is that cognition comprises computational operations defined over symbolic representations and that these computational operations are not tied to any given instantiation [48, 9, 49]. A physical body may facilitate exploration and learning, but it is by no means necessary. The principled decoupling of the cognitivist computational model of cognition from its instantiation as a physical system is referred to as *computational functionalism* [34]. The chief point of computational functionalism is that the physical realization of the computational model is inconsequential to the model: any physical platform that supports the performance of the required symbolic computations will suffice, be it computer or human brain. Computational functionalism effectively says that the mind is the software of the brain *or any functionally equivalent system.* This is an important claim:

[1] The literature on embodiment and embodied cognition is varied and extensive; see [49], Chapter 5, for a brief overview.

> "Computational functionalism entails that minds are multiply realizable, in the sense in which different tokens of the same type of computer program can run on different kinds of hardware. So if computational functionalism is correct, then ... mental programs can also be specified and studied independently of how they are implemented in the brain, in the same way in which one can investigate what programs are (or should be) run by digital computers without worrying about how they are physically implemented." [34]

There is an alternative school of thought in cognitive science that takes a very different view on this, arguing that cognitive systems are intrinsically embodied and embedded in the world around them, developing through real-time interaction with their environment. From the point of view of embodiment, the way the cognitive agent perceives the world — its space of possible perceptions — derives not from a pre-determined, i.e., purely objective, world, but rather from the actions in which the system can engage. In other words, it is the space of possible actions facilitated by and conditioned by the particular embodiment of the cognitive agent that determines how that cognitive agent perceives the world. Thus, the cognitive system constructs and develops its own understanding of the world in which it is embedded, i.e., its own agent-specific and body-specific knowledge of its world. This position is encapsulated in the *embodied cognition thesis*.

> "Many features of cognition are embodied in that they are deeply dependent upon characteristics of the physical body of an agent, such that the agent's beyond-the-brain body plays a significant causal role, or physically constitutive role, in that agent's cognitive processing." [53]

Underpinning embodied cognition is the assertion that perception and action are mutually dependent and that the dependency acts in both directions: action depends on perception (this, at least, raises no cause for objection), but perception also depends on action and, importantly, on the state of the agent's body (this is a little less obvious, but there is a large body of psychological and neuroscientific evidence to support it, e.g., [4, 10, 19, 36]). The mutual dependence of perception and action implies a dependence of cognition on the embodiment of the cognitive agent and the actions that embodiment enables. This has a far reaching consequence: agents with different type of body understand the world differently. The dependence of percepts, and associated concepts constructed through cognitive activity, on the specific form of embodiment is a fundamental cornerstone of embodied cognition and emergent cognitive systems, in general.

There are three hypotheses on embodiment associated with the embodied cognition thesis: the conceptualization hypothesis, the constitution hypothesis, and the replacement hypothesis [43].

The position that the physical morphology — the shape or form — and motor capabilities of a system has a direct bearing on the way the cognitive agent understands the world in which it is situated is sometimes referred to as the *conceptualization hypothesis*. That is, the characteristics of an agent's body determine the concepts an organism can acquire, and so agents with different type of body will understand the world differently.

The idea that the body (and possibly also the environment) plays a constitutive rather than a supportive role in cognitive processing, i.e., that the body is itself an integral part of cognition, is referred to as the *constitution hypothesis*. The claim made by the constitution hypothesis is stronger than that made by the conceptualization hypothesis. Cognition is not only influenced and biased by the characteristics and states of the agent's body, the body and its dynamics also augment the brain as an additional cognitive resource. In other words, the way the body is shaped and the way in which it moves help it accomplish the goals of cognition without having to depend on brain-centred neural processing.

There is a third claim sometimes made by proponents of embodied cognition: that because an agent's body is engaged in real-time interaction with its environment, the need for representations and representational processes is removed. This is referred to as the *replacement hypothesis*. The point of this hypothesis is that there is no need for the cognitive system to represent anything, computationally or otherwise, because all of the information it needs is already immediately accessible as a consequence of its sensorimotor interaction.

While the potential attractions of embodied cognition are numerous — the real-time situated coupling between the cognitive system and the environment, the possible removal of the need for symbolic representations, the embedded and grounded exploitation of the environment by the cognitive system to facilitate cognitive activity and off-load cognitive work and scaffold enhanced capabilities — the current capabilities of cognitivist systems are far more advanced. This is reflected in the state of embodied cognition that is sometimes referred to as a research program rather than a mature discipline. It is a plausible and, to many, a very compelling thesis, but, despite the fact that it is now accepted as a mainstream alternative to cognitivism, much remains to be done to establish it as an established science with well-understood engineering principles. In other words, it is not clear how the principles of embodiment should be manifest in a cognitive architecture. Also, embodied cognition entails that many aspects of procedural and declarative knowledge are agent-specific and cannot be directly shared with other cognitive agents.

On the other hand, in the cognitivist tradition, knowledge can be exchanged directly among different forms of cognitive agent, exactly because of it divorces cognition and cognitive architectures from the agent body, relying instead on the acquisition of the knowledge necessary to perform whatever task is necessary from whatever source is available. The cognitive architecture, then, is the fixed part of the cognitive model [24], which is completed by the addition of appropriate knowledge. The dilemma for the architect is that adopting a non-embodied cognitivist approach simplifies the task of designing the cognitive architecture but ignores considerable psychological and neuroscientific evidence of the role the body plays in cognition. Conversely, adopting an embodied stance does recognize this role, but it adds considerable complexity and the need to incorporate principles that are not yet fully developed into the design.

5 The Dilemma of Autonomy

The third dilemma concerns autonomy. To understand why autonomy presents a dilemma when designing a cognitive architecture, we need to be clear what we mean by autonomy. Unfortunately, that's easier said than done [7, 14] and one can identify more than twenty types of autonomy [49]. What is common to most interpretations is the idea that autonomy relates to the degree of self-determination of a system, i.e., the degree to which a system's behaviour — its goals and the manner in which it achieves them — is determined by the system itself and not its environment, including other agents [40]. Thus, an autonomous system is not controlled by some other agent, but is self-governing and self-regulating, selecting its goals, determining how best to achieve them, and then acting accordingly [16].

However, if an external agent cannot exert a causal influence on an autonomous cognitive system, how can one get it to do something useful? We want autonomy, but we also want some control over the cognitive system. This seems to present the architect with the dilemma of having to choose between control and autonomy. However, the choice is a little deeper than that. Mirroring the dilemma of fidelity and the need to choose between opting for the completeness and generality of natural cognitive systems or expediency when designing cognitive architectures for practical application, it is useful to distinguish between biological and robotic autonomy [54].

In robotics, it is common to distinguish between adjustable, shared, sliding, and subservient autonomy, all more or less equivalent terms that are suggestive of ways of qualifying the degree of autonomy and the relative involvement of a human with the cognitive system in carrying out tasks and pursuing goals. In these modes of autonomy, the system controls its own behaviour only to some extent, with the goals being determined by the human with which it is interacting [30]. In such cases, it is necessary for the cognitive architecture to accommodate this sharing: what information does the autonomous agent share with the user and on what basis does it decide whether or not it should be shared, for example [44]? The architect must still devolve to the cognitive system some power to make independent decisions and, in essence, all we have done is push the autonomy dilemma a little further down the line. The resolution of the dilemma hinges on the impact of those decisions, striking a balance between a human retaining control over the choice of superordinate goal and giving the system sufficient freedom to select strategies adaptively in order to meet these goals. A solution to this problem may lie in exploiting the information-theoretic concept of empowerment [39] in the design of the cognitive architecture.

For biological autonomy, we can differentiate between *behavioural autonomy* and *constitutive autonomy* [14, 3]. Behavioural autonomy is concerned with the extent to which the agent sets its own goals and its robustness and flexibility in achieving them as it interacts with the world around it, including other cognitive agents. Constitutive autonomy is concerned with the internal organization and the organizational processes that keep the system viable, maintaining itself as an identifiable autonomous entity. Indeed, Maturana and Varela, whose work provided the inspiration for the enactive view of cognition, define autonomy as "the condition of subordinating all changes to the maintenance of the organization" [29].

Constitutive and behavioural autonomy are related: an agent cannot deal with uncertainty and danger if it is not organizationally equipped to do so. Behaviour depends on internal preparedness, but appropriate behaviour is also needed to allow the agent to bring about the necessary environmental conditions for constitutive autonomy to be able to operate effectively. This complementarity of the constitutive and the behavioural reflects two different sides of the characteristic of recursive self-maintenant systems [6] to deploy different processes of self-maintenance depending on environmental conditions, with constitutive and behavioural autonomy corresponding to the internal and external aspects of that adaptive capacity, respectively.

The dilemma is now whether to base the design of the cognitive architecture on organizational principles that are not overtly concerned with achieving goals as perceived by external agents, or to focus on behaviour, but perhaps at the cost of missing some key aspect of cognition, e.g., *homeostasis* [8, 5], with the autonomy of an agent being effected through a hierarchy of homeostatic self-regulatory processes [31, 55], similar to Damasio's hierarchy of levels of homeostatic regulation [11].

If the processes that support constitutive autonomy were also to give rise to behavioural autonomy, the dilemma might be resolved without compromise. Recent work proposing that constitutive autonomy derives from self-organization based on continual predictive inference of the causes of sensory perturbations, coupled with continual adaption by updating the prediction model *and* responding with actions that minimize the long-term average surprisal, suggests this might just be the case [41, 42, 50].

6 Conclusion

Before concluding, let us recap the three dilemmas. The dilemma of fidelity involves choosing between a general and complete cognitive architecture that is a faithful model of human cognition, derived from desiderata, and an architecture that is specific to a particular application domain, derived from use cases. In the former case, all of the relevant elements will be addressed and it will be a well-founded framework, but some elements may be included that are not relevant for a given application domain and the architecture may be more complicated than necessary. In the latter case, the architecture will be focused on and driven by user requirements, but it may not be well-grounded in theory and may miss key principles that underpin cognition. Also, it is unlikely to yield insights into a unified theory of cognition.

The dilemma of embodiment involves choosing between cognitivism and computational functionalism in which the agent's body plays no role in the cognitive process and an alternative paradigm in which, to a greater or lesser degree, the body does play a causal role. In the former case, adopting a non-embodied cognitivist approach simplifies the task of designing the cognitive architecture but ignores considerable psychological and neuroscientific evidence of the role the body plays in cognition. Conversely, adopting an embodied stance does recognize this role, but it adds considerable complexity and the need to incorporate design principles that are not yet fully developed.

The dilemma of autonomy involves choosing between independence and control. Choosing independence, even partial independence and shared autonomy, creates the problem of how to incorporate the required restrictions on freedom to act into the cognitive architecture. On the other hand, choosing control simplifies the design of the cognitive architecture but undermines one of the key aspects of cognition: autonomous operation. If the goal is to emulate biological autonomy, the dilemma involves choosing between constitutive autonomy, focussing on organizational principles that are not overtly concerned with achieving the goal of external agents, and behavioural autonomy, with the potential of missing some key organizational aspect of cognition.

To conclude, let us look at the position that the editors of this book adopt on cognition:

> "In what concerns living systems, cognition is an embodied, embedded and always situated experience. This means it involves a cognitive entity endowed with a particular physical architecture interacting with the specific world it is immersed in, behaving according to the prompts placed by this environment, reacting, adapting to it, and this way defining its own existential narrative and history."

It is apparent that the editors have already confronted the dilemmas identified in this chapter and clearly favour the choices that see a cognitive system as a self-organizing biologically-plausible entity exhibiting complete autonomy, embodied and focussed on development. Furthermore, they add the following:

> "Highlighting the nature of the dialectics that binds different life forms to their specific environments, the book addresses the topic of artificial cognition in the domains of robotics and artificial life."

The key word here is *dialectics*, suggesting a process of never-ending discovery by which mutual interactions continually reveal new depths of meaning, at least insofar as the relationship between the agent and its world is concerned. This is the very essence of the concept of co-determination: the mutual specification of the system's reality by the system and its environment [28], strongly echoing the links between cognition, embodiment, and constitutive autonomy, succinctly captured by Anil Seth: "the purpose of cognition (including perception and action) is to maintain the homeostasis of essential variables and of internal organization ... [so that] ... perception emerges as a *consequence* of a more fundamental imperative towards organizational homeostasis, and not as a stage in some process of internal world-model construction" [42].

Having addressed the dilemmas and decided which choices best match the cognitive architecture design goals, the architect still faces a daunting challenge, but at least some of the design decisions are explicitly laid bare.

References

[1] Anderson JR (1996) Act: A simple theory of complex cognition. American Psychologist 51:355–365

[2] Anderson JR, Bothell D, Byrne MD, Douglass S, Lebiere C, Qin Y (2004) An integrated theory of the mind. Psychological Review 111(4):1036–1060

[3] Barandiaran X, Moreno A (2008) Adaptivity: From metabolism to behavior. Adaptive Behavior 16(5):325–344

[4] Barsalou LW, Niedenthal PM, Barbey A, Ruppert J (2003) Social embodiment. In: Ross B (ed) The Psychology of Learning and Motivation, vol 43, Academic Press, San Diego, pp 43–92

[5] Bernard C (1878) Leçons sur les phénomènes de la vie commun aux animaux et végétaux. J.-B. Baillière, Paris

[6] Bickhard MH (2000) Autonomy, function, and representation. Communication and Control—Artificial Intelligence 17(3–4):111–131

[7] Boden MA (2008) Autonomy: What is it? BioSystems 91:305–308

[8] Cannon WB (1929) Organization of physiological homeostasis. Physiological Reviews 9:399–431

[9] Clark A (2001) Mindware – An Introduction to the Philosophy of Cognitive Science. Oxford University Press, New York

[10] Craighero L, Fadiga L, Rizzolatti G, Umiltà CA (1999) Movement for perception: a motor-visual attentional effect. Journal of Experimental Psychology: Human Perception and Performance 25(6):1673–1692

[11] Damasio AR (2003) Looking for Spinoza: Joy, sorrow and the feeling brain. Harcourt, Orlando, Florida

[12] Dickmanns E (2003) A general cognitive system architecture based on dynamic vision for motion control. Journal of Systemics, Cybernetics and Informatics 1(5):1–6

[13] Franklin S, Madl T, D'Mello S, Snaider J (2014) Lida: A systems-level architecture for cognition, emotion, and learning. IEEE Transactions on Autonomous Mental Development 6(1):19–41

[14] Froese T, Virgo N, Izquierdo E (2007) Autonomy: a review and a reappraisal. In: Almeida e Costa F, Rocha L, Costa E, Harvey I, Coutinho A (eds) Proceedings of the 9th European Conference on Artificial Life: Advances in Artificial Life, Springer. doi: 10.1007/978-3-540-74913-4_46, Berlin Heidelberg, vol 4648, pp 455–465

[15] Gomez Esteban P, Cao H, De Beir A, Van De Perre G, Lefeber D, Vanderborght B (2016) A multilayer reactive system for robots interacting with children with autism. In: Proceedings of the Fifth International Symposium on New Frontiers in Human-Robot Interaction

[16] Haselager WFG (2005) Robotics, philosophy and the problems of autonomy. Pragmatics & Cognition 13:515–532

[17] Krichmar JL (2012) Design principles for biologically inspired cognitive architectures. Biologically Inspired Cognitive Architectures 1:73–81

[18] Krichmar JL, Edelman GM (2006) Principles underlying the construction of brain-based devices. In: Kovacs T, Marshall JAR (eds) Proceedings of AISB '06 - Adaptation in Artificial and Biological Systems, University of Bristol, Bristol, Symposium on Grand Challenge 5: Architecture of Brain and Mind, vol 2, pp 37–42

[19] Lackner JR (1988) Some proprioceptive influences on the perceptual representation of body shape and orientation. Brain 111:281–297

[20] Laird JE (2008) Extending the soar cognitive architecture. In: Proceedings of the First Conference on Artificial General Intelligence, IOS Press, Amsterdam, The Netherlands, pp 224–235

[21] Laird JE (2009) Towards cognitive robotics. In: Gerhart GR, Gage DW, Shoemaker CM (eds) Proceedings of the SPIE — Unmanned Systems Technology XI, vol 7332, pp 73,320Z–73,320Z–11

[22] Laird JE (2012) The Soar Cognitive Architecture. MIT Press, Cambridge, MA

[23] Laird JE, Newell A, Rosenbloom PS (1987) Soar: an architecture for general intelligence. Artificial Intelligence 33(1–64)

[24] Laird JE, Lebiere C, Rosenbloom PS (2017) A standard model of the mind: Toward a common computational framework across artificial intelligence, cognitive science, neuroscience, and robotics. AI Magazine In Press

[25] Lieto A, Bhatt M, Oltramari A, Vernon D (2017) The role of cognitive architectures in general artificial intelligence. Cognitive Systems Research in press

[26] Marr D (1982) Vision. Freeman, San Francisco

[27] Marr D, Poggio T (1977) From understanding computation to understanding neural circuitry. In: Poppel E, Held R, Dowling JE (eds) Neuronal Mechanisms in Visual Perception, Neurosciences Research Program Bulletin, vol 15, pp 470–488

[28] Maturana H, Varela F (1987) The Tree of Knowledge — The Biological Roots of Human Understanding. New Science Library, Boston & London

[29] Maturana HR, Varela FJ (1980) Autopoiesis and Cognition — The Realization of the Living. Boston Studies on the Philosophy of Science, D. Reidel Publishing Company, Dordrecht, Holland

[30] Meystel A (2000) From the white paper explaining the goals of the workshop: "Measuring performance and intelligence of systems with autonomy: Metrics for intelligence of constructed systems". In: Messina E, Meystel A (eds) Proceedings of the 2000 PerMIS Workshop, NIST, Gaithersburg, MD, U.S.A., vol Special Publication 970

[31] Morse A, Lowe R, Ziemke T (2008) Towards an enactive cognitive architecture. In: Proceedings of the First International Conference on Cognitive Systems, Karlsruhe, Germany

[32] Newell A (1982) The knowledge level. Artificial Intelligence 18(1):87–127

[33] Newell A (1990) Unified Theories of Cognition. Harvard University Press, Cambridge MA

[34] Piccinini G (2010) The mind as neural software? Understanding functionalism, computationalism, and computational functionalism. Philosophy and Phenomenological Research 81(2):269ñ–311

[35] Ramamurthy U, Baars B, D'Mello SK, Franklin S (2006) LIDA: A working model of cognition. In: Fum D, Missier FD, Stocco A (eds) Proceedings of the 7th International Conference on Cognitive Modeling, pp 244–249

[36] Rizzolatti G, Craighero L (2004) The mirror neuron system. Annual Review of Physiology 27:169–192

[37] Rosenbloom P, Laird J, Newell A (eds) (1993) The Soar Papers: Research on Integrated Intelligence. MIT Press, Cambridge, Massachusetts

[38] Rosenbloom PS, Demski A, Ustun V (2016) The sigma cognitive architecture and system: Towards functionally elegant grand unification. Journal of Artificial General Intelligence 7:1–103

[39] Salge C, Polani D (2017) Empowerment as a replacement for the three laws of robotics. Frontiers in Robotics and AI 4

[40] Seth A (2010) Measuring autonomy and emergence via Granger causality. Artificial Life 16(2):179–196

[41] Seth AK (2013) Interoceptive inference, emotion, and the embodied self. Trends in Cognitive Sciences 17(11):565–573

[42] Seth AK (2015) The cybernetic Bayesian brain — from interoceptive inference to sensorimotor contingencies. In: Metzinger T, Windt JM (eds) Open MIND, vol 35, Frankfurt am Main: MIND Group, pp 1–24

[43] Shapiro L (2011) Embodied Cognition. Routledge

[44] Sheridan TB, Verplank WL (1978) Human and computer control for undersea teleoperators. Tech. rep., MIT Man-Machine Systems Laboratory

[45] Sun R (2004) Desiderata for cognitive architectures. Philosophical Psychology 17(3):341–373

[46] Sun R (2007) The importance of cognitive architectures: an analysis based on clarion. Journal of Experimental & Theoretical Artificial Intelligence 19(2):159–193

[47] Sun R (2016) Anatomy of the Mind: Exploring Psychological Mechanisms and Processes with the Clarion Cognitive Architecture. Oxford University Press

[48] Varela FJ (1992) Whence perceptual meaning? A cartography of current ideas. In: Varela FJ, Dupuy JP (eds) Understanding Origins – Contemporary Views on the Origin of Life, Mind and Society, Kluwer Academic Publishers, Dordrecht, Boston Studies in the Philosophy of Science, pp 235–263

[49] Vernon D (2014) Artificial Cognitive Systems — A Primer. MIT Press, Cambridge, MA

[50] Vernon D (2016) Reconciling constitutive and behavioural autonomy: The challenge of modelling development in enactive cognition. Intellectica: The Journal of the French Association for Cognitive Research 65:63–79

[51] Vernon D (2017) Two ways (not) to design a cognitive architecture. In: Chrisley R, Müller VC, Sandamirskaya Y, Vincze M (eds) Proceedings of EUCognition 2016, Cognitive Robot Architectures, European Society for Cognitive Systems, Vienna, vol CEUR-WS Vol-1855, pp 42–43

[52] Vernon D, von Hofsten C, Fadiga L (2016) Desiderata for developmental cognitive architectures. Biologically Inspired Cognitive Architectures 18:116–127

[53] Wilson RA, Foglia L (2011) Embodied cognition. In: Zalta EN (ed) The Stanford Encyclopedia of Philosophy

[54] Ziemke T (2008) On the role of emotion in biological and robotic autonomy. BioSystems 91:401–408

[55] Ziemke T, Lowe R (2009) On the role of emotion in embodied cognitive architectures: From organisms to robots. Cognition and Computation 1:104–117