Epigenetic Robotics Architecture (ERA)

Anthony F. Morse, Joachim de Greeff, Tony Belpeame, and Angelo Cangelosi

***Abstract*—In this paper, we discuss the requirements of cognitive architectures for epigenetic robotics, and highlight the wider role that they can play in the development of the cognitive sciences. We discuss the ambitious goals of ongoing development, scalability, concept use and transparency, and introduce the epigenetic robotics architecture (ERA) as a framework guiding modeling efforts. A formal implementation is provided, demonstrated, and discussed in terms of meeting these goals. Extensions of the architecture are also introduced and we show how the dynamics of resulting models can transparently account for a wide range of psychological phenomena, without task dependant tuning, thereby making progress in all of the goal areas we highlight.**

***Index Terms*—Cognitive robotics architecture, concept use, conceptual learning through development, ongoing development, scalable and generality, transparent modeling, using robots to study development and learning.**

I. Introduction

INCREASINGLY, researchers across the cognitive sciences are calling for an approach to modeling that can scale up beyond simple scenarios, one that is not tailored to specific domains and tasks, displays an ongoing developmental trajectory, and is transparent in its creation and use of concepts [1]–[5]. While many cognitive architectures aim to achieve some combination of these goals, they too often provide only abstract frameworks, either lacking sufficient detail for implementation, or where implementation details are provided, the complexity of the resulting system is intimidating. Despite such complexity, to date relatively few approaches to modeling have displayed much success beyond any one of these aims. Arguably, such complexity may be a necessary feature of generally cognitive systems with wide ranging abilities; certainly in robotics, the combination of supporting systems required for such things as walking, manipulating objects, vision, and so on introduce a great deal of complexity before the issue of cognition and conceptual thought is even introduced. Despite this complexity and heterogeneity, as we demonstrate herein, significant progress can be made toward these goals of generality, scalability, development, and transparency within a simple homogeneous architecture based on intuitive dynamical principles of operation. The epigenetic robotics architecture (ERA) introduced here, provides such a homogeneous and intuitive framework that is simple to implement, and easily scaled and extended while making significant progress toward achieving all of these goals. Before providing the details of our proposed architecture, we first discuss a series of issues highlighting what it is that we believe an architecture should do for current cognitive robotics and cognitive science more generally.

A. What Are Architectures For?

From the very beginning, the aims and goals of AI have been hugely ambitious; to simulate every aspect of learning or any other feature of intelligence. Among those at the famous Dartmouth conference in 1956, often cited as the birth of AI, was Allen Newell. Newell's concern was that psychology could not mature by the accumulation of experimental data alone. As he wrote years later, "Suppose that in the next thirty years we continued as we are now going. Another hundred phenomena, give or take a few dozen, will have been discovered and explored. Another forty oppositions will have been posited and their resolution initiated. Will psychology then have come of age?" [4]. Clearly, more than 30 years later, the simple answer is no. Before psychology and equally, cognitive science can mature, the accumulation of scientific knowledge must fit, support, form, or revise wider theoretical perspectives in the quest for "grand" theories and/or better paradigms for understanding cognition.

While many researchers in the cognitive sciences are conducting hugely important experiments and developing crucial models exposing and accounting for various aspects of cognition, life, development, and so on, all of this takes place within one overarching theoretical paradigm. The primary role of an architecture as we see it, is to bridge this gap between theory and experiments or models in a manner (at least partially) formalizing how they can subsequently be integrated. This is not to suggest that models and experiments do not already fit the larger theories, but rather that more often than not various models and experiments which all make strong connections to the same theory do not seem entirely compatible themselves. To continue with Newell's example "...our task in psychology is first to discover that structure which is fixed and invariant so that we can theoretically infer the [collection of] method[s]" [4]. That is to suggest, not that we replicate human biocognition in every detail, for what would that tell us, but rather that we identify the underlying processes and how they are influenced, interconnected, and potentially interdependent. To this end, Newell proposed the modeling of *control* structures, which later became the field of cognitive modeling.

For Newell, cognitive modeling was intended to influence the development of psychological theory in several specific ways,

Manuscript received March 01, 2010; revised September 01, 2010; accepted October 04, 2010. Date of publication October 14, 2010; date of current version December 10, 2010. This work was supported in part by a European grant to the project "Integration and Transfer of Action and Language Knowledge in Robotics" (ITALK) Project 214668, and in part by another European grant to the project "Integrating Cognition Emotion and Autonomy" (ICEA) Project 027819.

The authors are with the Centre for Robotics and Neural Systems, University of Plymouth, Plymouth, Devon, PL4 8AA, U.K. (e-mail: anthony.morse@plymouth.ac.uk; joachim.degreeff@plymouth.ac.uk; tony.belpeame@plymouth.ac.uk; angelo.cangelosi@plymouth.ac.uk).

Color versions of one of more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TAMD.2010.2087020

which we believe are equally true for cognitive architectures: first, by providing a working model of some set of phenomena, the theorist is forced to be explicit about the specific details required to make that account work, thus an implemented simulation is far more rigorous and theoretically tight than an abstract flow chart, or some other boxology. Such implemented simulation also provides a working demonstration that the account given can truly produce the set of phenomena it is supposed to explain; the testability of the theory is guaranteed. Second, Newell suggested that modelers "...accept a single task and model all of it" [4] which is to say that all aspects of the phenomenon in question are explained and produced by the same model. With respect to Newell's first and second suggestions, current cognitive modeling has been a huge success, but this restricts the generality of the resulting models by accounting only for specific tasks. Newell's third suggestion, "to stay with the diverse collection of small experimental tasks, as now, but to construct a single system to perform them all" [4] presents the greater challenge facing modern cognitive modelers and would-be designers of cognitive architectures.

Despite the huge influence that the various computational approaches to modeling have had in the cognitive sciences (for example, the commonplace use of computational language), truly integrated cognitive modeling is extremely rare, even in modern robotics. And though Newell's views were expressed more than 30 years ago, and a great deal of progress has been made since then, the essence of his critique remains highly relevant to today's modelers, providing a rational underling our need for cognitive architectures. In summary, architectures for cognitive robotics should (at least attempt to) account for the integration of various cognitive phenomena and models from wide ranging domains into a single unified system, while providing sufficient guidelines for implementation and subsequent testing.

B. Why Bother With Autonomous Robots?

The use of robots forces modelers to address the integration of cognitive features all the way from the sensory surface to the motor surface of a robot, which while not forcing domain generality, clearly provides a contrast to traditionally disembodied and isolated approaches in which models are provided with task relevant representations of the state of the environment, thereby avoiding the modeling of perception and introducing both design bias and grounding problems [6], [7]. In robotics, any model that cannot span this gap is, quite simply, insufficient to be the controller of an autonomous robot and must be rejected. While autonomous robotics, as a domain of research, clearly does not force the modeling and integration of all cognitive phenomena, it is interesting that the inclusion of, even simple, artificial embodiment forces the integration of perception, thought, and action, and represents a significant step toward the truly unified systems required to provide the integration that cognitive scientists ultimately require of their theories. This would seem not only to be a tool for better understanding human cognition, but also for understanding and developing cognitive systems more generally.

Of course, the development of cognitive robots is, for many researchers, a goal in itself, but the radical changes in the ways that we design and model cognitive systems, resulting from cognitive robotic enterprises, have wider implications for understanding cognition. It should also be obvious that such integration of the facets of cognition, whether through cognitive robotics or cognitive modeling, does not need to have a lot to do with embodiment or sensorimotor approaches. Indeed, there are various notable approaches to the integration and unification of computational models. Perhaps the most well-known example is the subsumption architecture [8] in which complex behaviors are decomposed into layers of behavioral modules, each implementing successively more complex goals while subsuming the decisions of previous layers. Thus, the decisions of low levels producing obstacle avoidance can be influential in higher layers, producing goal-directed movements. Such approaches to behavior-based robotics follow a different decomposition to more traditional Sense-Model-Plan-Act decompositions, thereby solving engineering problems and hinting at potential biological solutions. As with most cognitive architectures, the complexity of implementation quickly becomes intimidating and difficult to manage. Though subsumption does not necessarily have to be achieved computationally or even hand designed, in practice it typically is.

In another branch of embodied modeling following a dynamical systems approach [9], [10], agents are typically controlled by neural networks evolved to solve minimally cognitive tasks. Here, the complexity of integrating perception, thought, and action is offset by the simplicity of the environment, embodiment, and task, and rarely is the resulting control system considered in traditionally cognitive terms. Nevertheless, such approaches consistently provide surprising insights into both the biological and psychological processes underlying cognition. Unfortunately, with current methods the search/design problem scales exponentially with the complexity of desired behavior of the resulting agent, which along with the increasing difficulty of evaluating the fitness of phenotypes, is the primary reason why evolutionary models remain at the level of minimal cognition. Despite these problems, there are some promising approaches attempting to resolve these issues [11].

From a dynamical systems perspective, embodiment itself is equally crucial to understanding cognitive processes, which emerge from the coupling between agent and environment. For example, though bipedal robots can be made to walk using largely disembodied methods, such as zero moment point, careful consideration of the human body (and bodies of other animals) reveals an ability to walk in a narrow range of conditions purely based on passive dynamics. This narrow niche can then be extended through the addition of powered movements working with and influencing the forces exerted on the body's natural dynamics [12]. Such powered passive dynamics approaches require a fraction of the computational power of more traditional methods for nondynamic bodies such as zero moment point. They also provide similar energy consumption during walking to their human counterparts, and result in far more natural walking gates [12], [13]. The point here is to highlight the many and varied roles that embodiment can play, and that seemingly (computationally) complex phenomena can often be offloaded to some large extent into the body and the environment, massively simplifying the computational control

structures and processes required. Such paradigms clearly have much to offer, providing a radically different interpretation of cognition from the traditional computer–mind metaphor.

Taking advantage of this requires careful consideration of the role of the body and the environment in extending our cognitive abilities [1], [14]. Moreover, careful consideration, not only of embodiment and autonomy, but also of the interdependence of cognitive functions leads to a reappraisal of the boundaries of cognition. For example, areas traditionally treated separately such as action production, skills, scene analysis, and representation are, following these developments, better considered as inseparably interdependent and fall under the more general label of perception. These varied reasons all support the important role that cognitive robotics research can play in the development of cognitive architectures.

II. What Should an Architecture Do?

Having both summarized the highly ambitious motivations for designing cognitive architectures and emphasized some of the important contributions that cognitive robotics can make, we now turn to the question of exactly what we want to achieve with our cognitive architecture. Clearly, a fully transparent model of all cognition (if such a thing is even possible) would be fantastic, but in making steps toward that ultimate goal, we are currently setting our sights a little lower. In light of the arguments that we have reviewed in the introduction, this section now defines three aspects of cognition that are central to cognitive architectures in epigenetic robotics.

A. Ongoing Development

In developmental psychology and epigenetic robotics, it is clear that cognitive systems should not be viewed as static systems that, once designed, are to be performing perfectly in their designated niche, but rather should undergo an ongoing development through interaction with their environment. Cognition, at least in any natural form, is never the result of genetics alone [15], [16], but results from the interplay of genetics (or in the case of nonevolved artificial systems, algorithm design) and the environment such that new behaviors, skills, and abilities emerge throughout the lifetime of the system. In many cases, this ongoing development follows a necessary sequence where the acquisition of certain abilities is a prerequisite of subsequently emergent ones. modeling cognitive systems that display an ongoing emergence or developmental path is proving highly challenging and progress in this important aspect of cognition has to date been extremely slow. Nevertheless, we believe that the importance of ongoing development should be a high consideration in the design of any cognitive architecture.

B. Concepts and Transparency

Following the aims of cognitive architectures set out in the introduction, cognitive architectures must account in some transparent way for the existence of the resulting cognitive behaviors and capacities. As we have already indicated, a fully detailed model of human biology would remain a biological theory, and on its own, would not necessarily advance our understanding of human cognition. Having said that, insights and abstractions from biology are extremely important, but equally important in developing an understanding of cognition is that there is some transparency in the principles of operation in a model or architecture. In even partially accounting for cognition, without transparency it would remain unclear that our models really are cognitive rather than simply appearing so. To this end, we stress that the hallmark of human cognition is the ability to develop, ground, manipulate, and otherwise use conceptual knowledge of the world. While we fully accept that cognitive architectures with alternative goals do not need to emphasize the importance of conceptual knowledge, for us this is a central requirement. By the same rational, it is not sufficient to simply assume and design-in the existence of concepts, symbolic, or otherwise, as this would not truly account for their existence. Rather, knowledge organized in recognizable conceptual structures should emerge from the system through interaction with its environment.

C. Scalability and Integration

Our final requirement is that a cognitive architecture should not be bound to a single domain of cognitive performance, but should integrate a wide range of phenomena and be scalable in its potential to integrate more. This would seem to constitute the major difference between a model and an architecture. In the next section, we now briefly review one theoretical paradigm having some potential to address the aspects of cognition that we have identified as central to the development of an epigenetic robotics architecture.

III. Constructivism and Sensorimotor Theories

Constructivist accounts of cognition assume that cognitive agents (human or otherwise) have an innate, but limited, and in some cases absent, knowledge of the world they inhabit, the actions that they can perform, and the likely effects of those actions. Extensions of this knowledge gained during the lifetime of an agent must then be constructed; developing from and remaining grounded in innate knowledge (primitives), the experiences that the agent has with its environment, and combinations thereof [6], [17]. Though typically focused on experience rather than innate knowledge, sensorimotor theories provide a highly intuitive account of the construction of world knowledge and other cognitive capacities during the lifetime of an agent. Such intuitive accounts also provide insights into the processes and mechanisms involved in biological cognition, question some of the well-established assumptions in cognitive science, and suggest a route to modeling the artificial development of cognitive systems in robotics, of which ERA is an example.

In considering both innate primitives and experience, the latter is viewed as a far richer source of information about an agent's current environment than innate primitives bestowed through evolution, which are presumably (but not necessarily) far less adaptive at the timescale of an individual. Despite this bias, it is important to view both cognitive and physical development as resulting from the interplay between an agent's genetic inheritance and its environmental experiences. Neither ontogeny nor cognition can result from nature or nurture alone and drawing such dichotomies can be highly misleading [15], [16]. Experiences then can never be the nature free experiences of an objective world that, as developed cognitive agents, we

are sometimes disposed to assume, rather for any agent, experiences are both embodiment and cognitive centric in that they are experiences of that agents bodily interactions with a world interpreted in the light of past experiences. They exist in the combination of sensory, somatic (available indictors of bodily well being, primarily; internal bodily sensations, metabolic factors, and hormonal influences), and motor factors. This is not to claim, as others have [18], that all knowledge has a motor component, nor that all knowledge is inextricably tied to the sensory, somatic and motor surfaces. To clarify, though all knowledge is clearly grounded in either these surfaces, innate primitives, or combinations thereof, much of the development and construction of cognition is a process of abstraction to the extent that all cognitive capacities reside on a continuum with these basic sources of information at one extreme [19]. Thus, proposing a continuum of abstraction rejects the necessity for qualitative distinctions and specialized mechanisms beyond those already implied in the most basic of sensorimotor learning systems. So while the developed human brain clearly makes use of a great deal of specialized circuitry, much of this circuitry is itself a product of the ongoing development of the agent in interaction with its environment [20], [21].

At the heart of all sensorimotor theories of cognition is the claim that perception is, to a large degree, based upon the use of sensorimotor knowledge in predicting the future sensory consequences of an action, either overtly executed or covertly simulated [3], [5], [22], [23]. From a modeling or architectural perspective, such an approach is appealing for two distinct reasons; first, as the sensory consequences of executed actions are readily available, most learning methods are applicable and there is no need for distinct training or learning phases. Second, sensorimotor accounts of cognition are inherently both modality and domain general in that they do not presuppose any specialized process specific to any particular modality or domain, rather the general process of sensorimotor learning is applied to whatever input/output streams happen to be available, and in whatever domain the information therein pertains to.

A. Perception

Our perception of continuous contact with a rich visual world laid out in front of us is somewhat misleading. In fact, our actual sensory input is highly impoverished. For example, visual acuity is focused on an area the size of a thumb nail at arm's length. From a sensorimotor perspective, our perception of objects outside the fovea (including those outside our current field of vision) is largely constructed from predictions of what you would see if were you to look in this or that direction. It is worth noting that such perception is clearly supported by processing of the sparse input from the periphery of our visual field, and mechanisms drawing attention to movement, flashes, and other changes and that these mechanisms are supportive of sensorimotor learning rather than inherently bound to it. As we move away from "low-level" sensorimotor predictions, objects should be identifiable through the profile of interactions that they afford. To use a common example; we can perceive a plate as round, not because it projects a round image onto our retina, but rather because we can predict how our sensory contact will change as we move a little this way or a little that way. This rather sparse account supposes that such profiles can be constructed and recognized leading to the recognition of objects in the world in terms of their Gibsonian affordances [24]. This construction of profiles of interaction is both crucial to the ability of sensorimotor theories to account for high level cognitive/mental perception, and is also the least detailed and most challenging aspect. Few sensorimotor theories do more than just suppose an ability to do this. Nevertheless, such embodiment centric accounts of perception are supported by a large number of psychology experiments exposing various bodily biases in categorization.

Barsalou *et al.* [25] highlights some of the ways in which body posture and action effect perception and cognition. For example, subjects rated cartoons differently when holding a pen between their lips than when holding it between their teeth. The latter triggered the same musculature as smiling, which made the subjects rate the cartoons as funnier, whereas holding the pen between the lips activated the same muscles as frowning and consequently had the opposite effect [26]. Moreover, bodily postures influence the subjects' affective state, e.g., subjects in an upright position experience more pride than subjects in a slumped position. Further compatibility between bodily and cognitive states enhances performance. For instance, several motor performance compatibility effects have been reported in experiments in which subjects responded faster to "positive" words (e.g., "love") than "negative" words (e.g., "hate") when asked to pull a lever towards them [27]. Smith and Samuelson [28] demonstrate that the typical spatial location of (and hence, body posture when reaching for) objects relative to the subject can be a stronger influence in learning their names than the correlation between hearing the new name and attending to a new object. We will return to Smith and Samuelson's experiments later in this paper, using them to demonstrate aspects of the implemented architecture.

IV. Conceptual Spaces

Having identified the transparent generation, grounding and use of conceptual structures as a key target of our architecture, we now briefly review the conceptual spaces theory of concepts. A classical distinction in most conceptual theories is whether concepts should be represented as theories themselves or as values in some parameter space. While the former method would be most suitable for a symbolic account of cognition, the latter would seem more compatible with connectionist and neural approaches. An interesting position in the middle is the so-called conceptual space (CS) as described by Gärdenfors [29]. Conceptual spaces are postulated as a way to represent knowledge on a level that resides in between the symbolic approaches on the one hand, and connectionist modeling on the other. Connectionist models and neural networks are seen as subconceptual, providing detailed processing of the lowest level of information units, while symbolic processing is seen as the most abstract form. To account for both the construction of concepts and their transparent use, our architecture must bridge this apparent methodological gap. As such, the CS is posed in between these levels, describing concepts in terms of geometrical shapes that are linked or grounded in sensory properties, but may also exhibit symbol-like behavior [19]. A CS consists of a

geometrical representation in vector space along various quality dimensions. In a nutshell, a CS is a collection of one or more domains (like color, shape, or tone), where a domain is postulated as a collection of inseparable sensory-based quality dimensions with a metric. Examples of quality dimensions are *weight*, *temperature*, *brightness*, *pitch*, *loudness*, and *RGB values*. For instance, to express a point in the color domain using RGB encoding, the different quality dimensions "red," "green," and "blue" are all necessary to express color values and are hence, inseparable. Other domains may consist of one or more quality dimensions. In its simplest form, a concept can be represented as a point in the conceptual space, where the coordinates of the point determine the features of the concept. For example, an instance of the concept RED may be represented as a point (255, 0, 0) in the RGB color domain and an instance of the concept BLUE as another point (0, 0, 255) in the same space. In principle, any domain may be used, although for some domains it might be easier to extract the relevant dimensions than for others.

A system equipped with conceptual knowledge structured in a CS can classify newly observed stimuli as belonging to a particular concept by calculating the weighted distance from the stimuli to every other instance of concepts already present in the CS. The observed stimulus is then assigned to the closest existing concept. Furthermore, a CS allows for the representation of concepts through prototypes, which enables it to display typicality effects observed in human conceptualization. Rosch [30] pointed out that many everyday concepts are prototypical in nature, i.e., humans regard certain instances for a specific concept to be more typical than others. For example, for the concept BIRD, the instance ROBIN is thought to be more "bird-like" than the instance PENGUIN. Hence, it seems that specific instances exhibit a graded membership to an idealized prototype. A conceptual prototype is built through the addition of exemplars for the specific concept, where the mean values of all quality dimensions encode for the coordinates of the prototype, and the variance of all exemplars determines the prototype's size. More general conceptual prototypes will therefore, due to their high variance, typically not be exact points in the CS, but rather define a certain convex region, while more specific prototypes may consist of an exact point in the CS. Through distance calculation from exemplar to prototype, it is possible to derive the most prominent exemplars for a certain concept, allowing the CS to exhibits both prototype and exemplar properties.

A. A Need for Richer Representations

Even though a CS could potentially be suitable for the representation of concepts, a number of drawbacks can be identified. First, it is not always clear from Gärdenfors work how the quality dimensions that are needed to express concepts can be found automatically as they are typically either defined beforehand or through multidimensional scaling. Second, the ability of CSs to represent nonlinear classes is highly limited. Gärdenfors suggests that nonlinearity can be dealt with by stretching dimensions through weighting depending on attention mechanisms, but this would clearly not work for many nonlinear classes. Even once proto-concepts are formed their subsequent use must be further explicated. Common use of concepts can be understood in terms of

- *Priming and Association*: the way humans use conceptual knowledge as part of their cognitive ability appears to be more elaborated than a simple matching of features process. For instance, priming (both perceptual and lexical) can greatly influence the way humans perceive and classify objects. So the manner in which an object is perceived depends not only on its perceptual features, but also on any priming that may have occurred before the observation. Also, an observation will typically not just activate one specific concept, but rather activate a "web" of associated concepts. For instance, when perceiving FIRE, the concept of FIRETRUCK, FIREMAN, and WATER may be readily available for most people, even though there is no clear perceptual similarity between these concepts.
- *Compositionality*: this refers to the ability to combine concepts into new concepts. This is not a simple conjunction of two concepts, where the meaning of the new one lays in between the other two. Rather, the first concept will act as a modifier on the second by imposing contextual limits and/or extensions on the features of the second concept. Sometimes this may be relatively straightforward, e.g., BIG CUP denotes a specific subset of all CUPs, namely the ones that are big. Others are notoriously harder, e.g., STONE LION (example given by Gärdenfors), because the modifier STONE creates a lot of properties that are not normally associated with LION, only the shape remains intact.
- *Hierarchy/Taxonomy*: concepts are typically part of a taxonomy, so knowledge about the properties of a concept higher up in the taxonomy tree generates access to knowledge about subordinate concepts. For instance, if it is known that a CAT is a subordinate of the concept MAMMAL, all sorts of properties may be inferred even though they do not need to be observed as belonging to CAT explicitly. In this case, we may know that MAMMAL is typically warm-blooded and feed their young, so this is inhered by the concept CAT. Such a hierarchical organization needs to be viewed separately from associative structures, as concepts may be associated with each other even though they are not in the same taxonomy (as is the case with the FIRE example given above)

V. The ERA

Following the extended discussion of the rather ambitious requirements that we have set out for a cognitive architecture, and several theoretical positions having relevance to meeting those varied requirements, we now introduce the core of ERA as an architectural framework and provide a simple example of implementation and use. On first appearances, the ERA architecture is rather simple, perhaps overly so consisting only of a homogeneous hierarchy, yet we believe the ERA framework to be; extremely powerful, consistent with constructivist sensorimotor theories, easily extended, and perhaps most importantly, to provide advances in scaling up beyond simple scenarios, in integration, cumulation and generality, and in displaying an ongoing developmental trajectory. We do not claim to have solved all of

the issues and developed an architecture that can do it all, rather our more modest claim is that we have made progress toward these ambitious goals as we shall discuss further at the end of this paper. We now turn to a more formal mode of modeling and describe the elements that constitute the ERA architecture.

A. The Basic ERA Unit: Structured Association Between Self-Organizing Maps

The basic unit of the ERA architecture is formed by the structured association of multiple self-organizing maps [31], resulting in structures with a strong resemblance to localist Interactive Activation and Competition (IAC) models which have a long history of use in modeling psychological phenomena [32]–[34]. Each self-organizing map (SOM) receives a subset of the input available to that unit and is typically partially prestabilized using random input distributed across the appropriate ranges for those inputs. For example, a SOM receiving three inputs as the average red, green, and blue values from a region of an image, can be prestabilized by training with randomly generated RGB input values such that it forms a conceptual color space. Increasing the probability of extreme values in the randomly generated training data ensures that the resulting stable map fully covers the range of possible input values, without this step mid-range values would tend to pull in the extremities of the map resulting in poor coverage of those extremes. Standard equations for generating SOMs are shown below.

Initial activation of SOM units

$$A_j = \sqrt{\sum_{i=0}^{i=n}(v_i - w_{ij})^2} \tag{1}$$

where A_j is the resulting activity of each node in the map following a forward pass, v_i is an input, and w_{ij} is the weight between that input and the current node. The winning node is the node with the smallest value for A_i.

Final activation of SOM units

$$y_i = e^{(\frac{-\beta_i}{2\sqrt{n}})} \tag{2}$$

where y_i is the final activation of the ith node in the map, β is the distance from node i to the winning unit, and n is the total number of nodes in the map. Note: units not within the neighborhood size are set to zero output activation, the neighborhood size and learning rate are monotonically decreased and the map is taken to be stable when the neighborhood size is zero.

Weight changes

$$\Delta w_{ij} = \alpha(v_i - w_{ij})y_i \tag{3}$$

where w_{ij} is the weight between input j and unit i, and α is the learning rate.

1) SOMs and CSs: A SOM bears considerable resemblance to a CS in the sense that both allow for clustering of multidimensional data. The formation of convex regions in CSs, serving as prototypes, may be seen as an analogue to the SOMs classification of topological locations encoding certain regions of the input data space. However, some differences can also be noted. These are, among others, the fact that SOMs are able to compress high dimensional data into a lower dimensional structure, from the input space to the SOM space, in effect economizing the representation. While such compression is important to the function of the ERA architecture as we shall see in the next section, to understand the functioning of the basic ERA unit it is useful to consider the weight space of the SOM rather than the lower dimensional SOM space and hence ignore the dimensionality reduction for now.

While data representation within a single domain (such as color) may yield similar results when using either CSs or SOMs, CSs tend to include dimensions from different domains (e.g., color, size, and texture), while SOMs in general tend to be applied only to one specific domain (though nothing in principle precludes their use in multiple domains). Even so, where different modalities drive different SOMs it is possible to use each SOM as a lower level representational structure incorporated on a higher level in a CS-like fashion with other SOM based lower level structures [29]. This would be somewhat analogous to the usage of special "hub" SOMs as we shall now demonstrate.

2) Putting SOMs Together: Many recent models of cognition have combined multiple SOMs together through association with great success, both in modeling the connection between different brain regions and in modeling psychological function. For example, Westermann and Miranda [35] demonstrate that associations between auditory and vocalization maps can be used to model "b"abbling leading to the emergence of vowel categories. Li *et al.* have demonstrated age of acquisition effects [36], and vocabulary spurts [37], following similar map-based modeling, while Mayor and Plunkett [38] demonstrate taxonomic responding where visual and auditory maps are associated. Caligiore *et al.* [39] demonstrate compatibility effects in biologically structured networks of maps, more closely related to the hierarchies discussed herein.

In the simplest case, the ERA architecture comprises of multiple SOMs, each receiving input from a different sensory modality, and each with a single winning unit. Each of these winning units is then associated to the winning unit of a special "hub" SOM using a bidirectional connection weighted with positive Hebbian learning using the following equation.

Positive Hebbian learning

$$\Delta w_{ij} = \alpha x_i x_j \tag{4}$$

where w_{ij} is the weight between node j and node i, α is the learning rate $(0.01), x_i$ is the activity of the winning node in one map, and x_j is the winning node in ahub map.

In some cases, one of the existing input SOMs can be selected as the hub but more often the hub SOM will provide pattern recognition over the activity of the other SOMs in the ERA unit (see Fig. 1). This can be achieved by taking the coordinates of the winning units (in SOM space) of each input SOM as an input to the "hub." Again, if we know the number of SOMs connecting to a "hub," then we can prestabilize that "hub" by training with random values in the appropriate ranges.

The absence of antihebbian learning leads to the possibility of certain SOM units having connections to many "hub" units

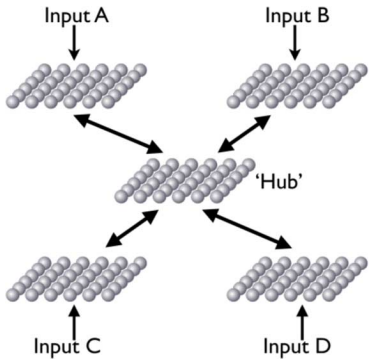

Fig. 1. Basic ERA unit. Multiple SOMs receive and classify different inputs, the winning units of which are associated to the winning unit of a "hub" SOM which either classifies the activity of the multiple input SOMs or classifies another external input.

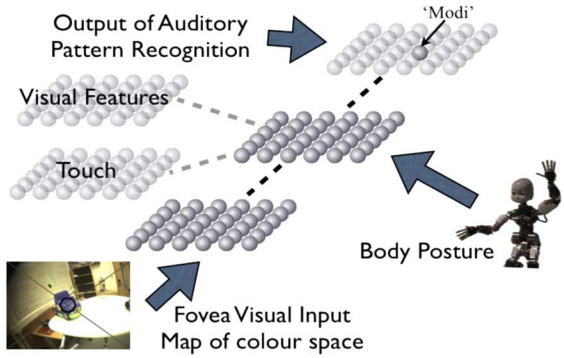

Fig. 2. General architecture of the model. SOMs are used to map the color space, the body posture, and the word space. These maps are then linked using Hebbian learning with the body posture map acting as a central "hub." The model can easily be extended to include other features such as visual and touch information in additional SOMs.

and vice versa, however, as associations are built up between "hub" units and units in the other SOMs, and as the "hub" itself is differentially sensitive to the context of activity over the other SOMs, this rarely happens in practice and is not detrimental to the functioning of the overall model. If the "hub"is not specialized and instead an existing input SOM is used as the "hub," then an abundance of Hebbian connections is far more likely to become a problem and methods for reducing the number of connections must be considered. Obvious examples to consider would include the normalization of connection strengths, the addition of antihebbian learning, or the decay of connections. Each of these methods have their own drawbacks and it is up to the modeler to make decisions appropriate to the domain of their modeling even though such decisions will ultimately limit the domain of application of the resulting models.

Having established the winning units in the various SOMs via a forward pass from the input activity, activity within the ERA unit then spreads via the bidirectional Hebbian connections in much the same way as an IAC network, allowing the presence of features or concepts in one map or CS to prime features or concepts in the other maps/conceptual spaces. In comparison to localist IAC models, each map can be considered to be equivalent to a pool of mutually inhibitory conceptual units and thus, the mutual inhibition and self-excitation already present in the SOM continues to operate. As with localist IAC models, the priming effect between SOMs/pools not only conforms to psychological priming [33], [34], but behaves as a *content addressable memory*, displaying *graceful degradation*, *default assignment*, *flexible generalization*, and crucially *emergent schemata* [40], [41]. Furthermore, such structures have, in explicitly localist incarnations, been shown to display a very wide range of psychological phenomena resulting from the same functional structure and so, in a very real sense, this structure or model seamlessly integrates those phenomena such that every such structure will display all of these phenomena in the domain of input used.

As Mike Page comments with reference to such localist models, "I make no claim to be the first to note each of these properties; nonetheless, I believe the power which in combination they afford has either gone unnoticed or has been widely underappreciated." [41]. We are in complete agreement with Page on this and further suggest that the ERA unit as described here relaxes the localist constraint of IAC models, and provides a grounded process of learning and development for these structures. As a simple demonstration of the basic unit of ERA in use, we now summarize the work published in Morse *et al.* [42]

3) The ERA Unit in Action: In a series of experiments related to Piaget's famous A-not-B error [43], and derived from experiments by Baldwin [44], Linda Smith, and Larissa Samuelson [28] repeatedly showed children between 18 and 24 months of age two different objects in turn, one consistently presented on the left, and the other consistently presented on the right. Following two presentations of each object, the child's attention is drawn to one of the now empty presentation locations and the linguistic label "modi" is presented. Finally, the children are presented with both objects in a new location and asked, "Can you find me the modi?" Not surprisingly, the majority (71%) of the children select the *spatially correlated* object despite the fact that the name was presented in the absence of either object. Varying the experiment to draw the child's attention to the left or right rather than to the specific location that the object, when saying "modi," resulted in a similar performance where 68% of the children selected the spatially linked object. The results of this experiment challenge the popular hypothesis that names are linked to the thing being attended to at the time the name is encountered.

The "modi" experiment, and its variations, strongly suggest that body posture is central to the linking of linguistic and visual information, especially as large changes in posture such as from sitting to standing disrupt the effect reducing performance in the first experiment to chance levels. In the basic ERA unit, we take this suggestion quite literally, using body posture information as a "hub," connecting information from other sensory streams in ongoing experience. Connecting information via a "hub" allows for the spreading of activation via this "hub" to prime information in one modality from information in another. Furthermore, using the body posture as a "hub" also makes a strong connection to sensorimotor theories of cognition; as actions, here interpreted as changes in body posture, also have the ability to directly rather than indirectly prime all the information associated with that new position and hence, indicate what the agent would expect to see were it to overtly move to that

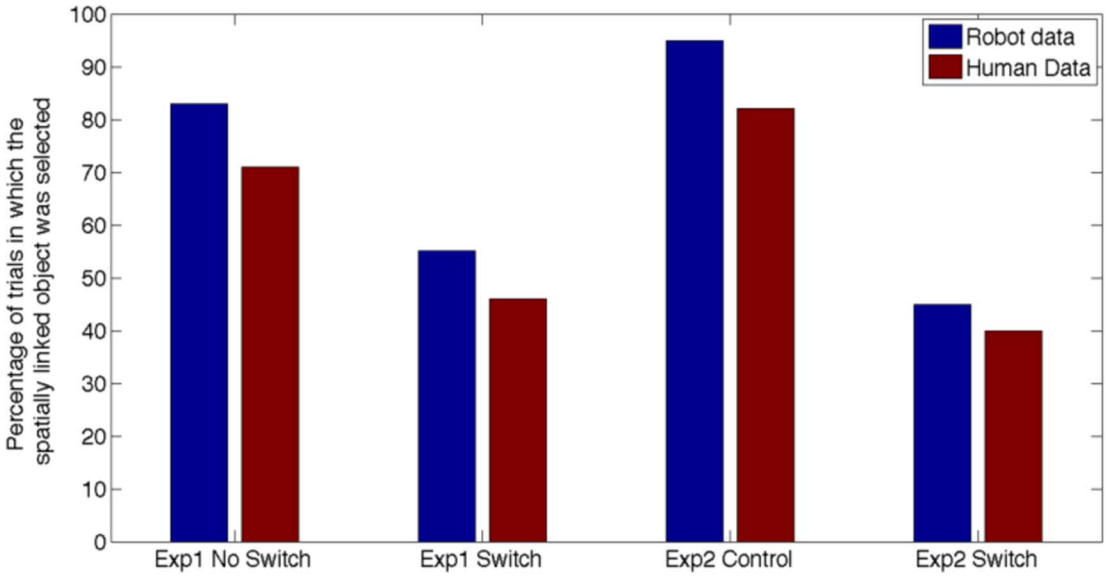

Fig. 3. Percentage of spatially linked objects selected in each experimental condition for both robot data and for the human–child data.

posture. As already discussed, such predictive abilities are the foundation of sensorimotor theories.

For this demonstration, we use the humanoid robotic platform iCub, an open source platform which has been recently developed as a benchmark platform for cognitive robotics experiments [45]. It has 53 degrees of freedom, allowing experiments on visual, tactile and proprioceptive perception, manipulation. and crawling. Initial iCub experiments were carried out in simulation through the open source iCub simulator [46], and then adapted and tested on the physical robot platform.

As the maps are linked together in real time based on the experiences of the robot (see Fig. 2) strong connections between objects typically encountered in particular spatial locations, and hence in similar body postures build up. Similarly, when the word "modi" is heard, it is also associated with the active body posture node at that time. Finally, at the end of the experiment, when the robot is asked to "find the modi," activity in the "modi" word node spreads to the associated posture and on to the color map node(s) associated with that posture. The result is to prime particular nodes in the color map, and the primed color is then used to filter the whole input image and the robot adjusts its posture to center its vision on the region of the image most closely matching this color.

The information linked via the body–posture hub is the result of processing visual input from the iCub robots cameras, taking the average RGB color of the foveal area and using this as an input to a 2-D SOM [31] described in (1), (2), and (3). The SOM provides pattern recognition over the input space preserving input topology while capturing the variance of the data. The body-posture "hub" similarly used the joint angles of the robot as input to another SOM. For simplicityherein, only two degrees from the head (up/down and left/right), and two degrees from the eyes (up/down and left/right) were actually used, thus the body map of the iCub robot has four inputs, each being the angle of a single joint. Finally, auditory input is abstracted as a collection of explicitly represented "words," each active only while hearing that word. In this demonstration, "words" are activated using the open source CMU Sphinx library (http://cmusphinx.org/) to provide voice processing. More detail on these experiments and a discussion of the results can be found in Morse *et al.* [42], but the main results (shown in Fig. 3) demonstrate a close fit between the data from the robot and that of children in the "modi" experiments.

By using the body posture as a central "hub" or orchestrator, the model predicts that while changes in posture (such as from sitting to standing) will disrupt the spatial naming effect, subsequent changes (such as moving back to sitting) will reinstate the effect. At the time of writing, these predictions are currently being tested in children. For comparison purposes, if the body posture was merely another input SOM and the "hub" was instead instantiated through simple SOM pattern recognition (a standard SOM with input as the x and y values of the winner in each input SOM), then changes in any input domain would be equally and weakly disruptive and the pattern of results shown in Fig. 4 would not be achieved. To give a clearer example, presentation of an object in a location is learned, however, the absence of the object and the presence of the spoken word is likely to result in a different "hub" unit winning, and so the connection between the spoken word and the visual object is lost.

B. The Full ERA Architecture: Hierarchies of ERA Units

Hierarchical structures are increasingly becoming popular in neural modeling, whether processing different levels of abstraction or operating with different time constants, examples can be found throughout the recent literature (e.g., [47]–[49]). While there is no requirement that cognitive robotics must mimic the underlying biology and neuroscience of human and animal cognition, the general structure of pathways is evident in the cerebral cortex of all vertebrates, all of which share the same basic brain organization [50]. While structures in the developing cortex vary significantly, Jones and Powell's [51] study of converging pathways highlights that information from each sensory modality, arriving in different regions of cortex, follows a similar path through the cortex. By implication, this

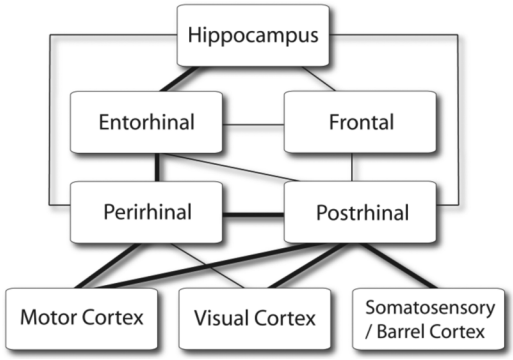

Fig. 4. Connection diagram showing the major pathways by which sensory information reaches specific regions of the rat cortex. The thickness of the connecting lines indicates the size of the projection.

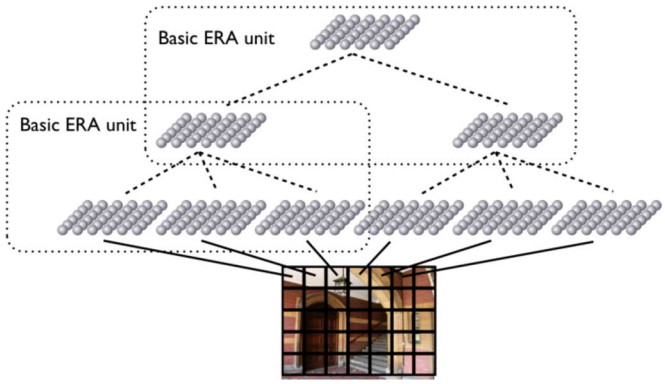

Fig. 5. Combining ERA units in hierarchies, a minimal example.

would suggest that different modalities, including the motor cortex, are treated in much the same way.

Thus, sensory information is projected into different regions of cortex where it progresses through several unimodal regions (such as the visual and somatosensory cortexes shown in Fig. 4) to various polymodal regions and then onto regions associated with motor function, and finally to the motor cortex itself. Brown and Aggleton [52] provide a more detailed map of the connectivity or flow of information via major pathways, in both the macaque monkey and rat cerebral cortex, a simplified version (including fewer sensory modalities) of which is shown in Fig. 4. Many such maps of the pathways between regions exist and all such maps can form the basis of modeling with the full ERA architecture, we simply selected this one as an example.

1) A Basic Hierarchy: Following on from the basic ERA unit, and roughly following Hawkins and Blakeslee. [48], Swanson [53], and Downing [2], [54], each microcolumn in a unimodal input region receives topographical input from a small area of whichever sensory modality targets that region. For example, each microcolumn in a rat's somatosensory cortex may receive input from a single whisker, and each microcolumn in the input region of the visual cortex (area V1) may receive input from a small area of the retinal image. To this end, every SOM in a particular ERA unit should now only receive a small number of inputs from neighboring areas of a specific modality. Many such ERA units may be required to fully cover the input stream from one modality. As microcolumns in any one region are not significantly interconnected (other than local inhibition), then the processing of sensory input in any one ERA unit is unable to function as a detector for features distributed more widely than its input. Thus in these regions, only relatively small, specific, and fast changing (due to the movement of the body), features can be detected (to use vision as an example, blobs, line segments, orientations, gradients, and so on).

Following the example of the visual cortex, in the human brain visual input targets area V1 which can be modelled as a large number of ERA units with some overlap of the regions of visual input to which they are responsive. Major pathways then connect area V1 to V2, combining the output of several cortical microcolumns in V1 into single microcolumns in V2. This is abstractly modeled in ERA by taking the "hub" SOMs from several ERA units in the input layer as input to "hub" SOMs in a new ERA unit in the next layer. Repeating this process through several layers allows high dimensional inputs such as images from a camera to be distributed across a large number of input layer SOMs and gradually combined through subsequent unimodal layers (see Fig. 5).

In practice, it is advantageous in unimodal regions to use the same SOM weights for each input SOM, as the same feature will then be identified as such despite changes in its position. Polymodal regions can be also constructed in much the same way by taking the output of "hubs" from ERA units operating in different modalities and combining them in new ERA units following known pathways between different brain regions, or even as a basic hierarchy ignoring biological pathways. Functionally, in the brain, each biological microcolumn or ERA unit detects and classifies features in its input, passing these feature classifications onto the next region [48]. Central to the functioning of both the real cortex and the ERA model is that while classifications of detected features flow up this hierarchy, top–down connections also project back along these pathways such that partial patterns are completed top–down providing anticipatory input based on the presence of other sensory features. This mechanism is essential to the architecture put forward here. As pathways from different sensory regions converge in polymodal regions, these regions, or ERA units, are able not only to detect multimodal features, but also to predict features in one modality based on information from another.

As an example of this bottom–up and top–down structure, though far from being a sensorimotor model, a popular and relatively successful approach to object recognition in computer vision is to generate a hierarchy of increasingly complex features; at the initial level we may find very low level features such as orientation and line detection, derived from raw image data. At the next level, these features can be combined into more complex organizations such as corners and edges, and at the next, perhaps shapes like curves, squares, and so on. While features at each level can be learned in a bottom–up way from experience of the coactivation of features in the previous layer, such hierarchies also allow for top–down projection, meaning that bottom–up partial evidence for a feature, sufficiently activating that feature leads to a top–down projection, from that feature, activating or priming the missing features at the previous levels. If this process of bottom–up and top–down spreading activation

is managed in an interactive activation and competition (IAC) manner [55], [56] as is the case in the ERA model then the process can be highly successful at visual scene based object recognition [57], [58]. There is a strong analogy between the workings of the ERA architecture and this kind of feature hierarchy in computer vision, though the methods of its implementation differ considerably. To avoid any confusion where we refer to a hierarchy, it is of this kind and we do not mean to imply any kind of executive control hierarchy.

This concludes the description of the ERA architecture. We now turn to a discussion of its relation to the theories introduced in the introduction, a discussion of possible extensions, and further examples of its use in cognitive robotics.

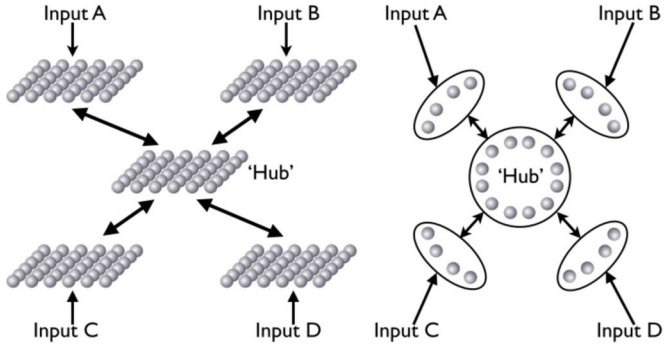

Fig. 6. Comparison between the basic ERA unit (left) and a localist IAC model (right). In the IAC model each node has a localist interpretation and inhibitory connections to other nodes in the same pool (represented by circles).

VI. ERA in Relation to Theory

The ERA architecture whether interpreted as a set of guidelines for integration or as a specific modeling paradigm makes strong links between various theories in cognitive science and a level of modeling, especially robotic modeling, appropriate to their instantiation and integration. We begin with the relation to theory.

A. ERA and Constructivist Sensorimotor Theories

The design of the ERA architecture is the result of a long lineage of models specifically aimed at instantiating theories of sensorimotor perception [3], [5], [22], [23]. At the heart of these theories is the ability to predict the future sensory consequences of actions, which ERA allows for by the spreading of activation between connected subhierarchies of sensory and motor modalities. When actions are performed they are associated with the sensory input at that time, thus any consistent consequences on sensory input become strongly associated to those motor actions. By this mechanism the covert simulation of motor actions will result in the priming of the predicted consequences of those actions. Even in a nonhierarchical form, such as the "modi" experiment discussed previously, activation in potential motor areas or in the resulting body posture acting as a hub can directly stimulate, via a spreading of activation, sensory regions as predictions of the sensory consequences of actually performing those actions. Similarly perception or recognition of any object will, again via a spreading of activation, prime specific motor areas and can itself be disrupted by competing activity in the motor cortex (resulting in primed competition). Such motor activity resulting from perception is completely compatible with the significant body of evidence of activity in the motor cortex being part of even nonmotoric perceptions.

1) The Importance of Context: While it is fairly easy to see how the ERA architecture can predict the sensory consequences of actions in simple and contrived scenarios, a more complex account of the dynamics of ERA models is required for a more general account of sensorimotor perception. The first step is to consider the strong analogy between the basic ERA unit and connectionist IAC models. The mutual inhibition implied in the winner-takes-all execution of each SOM is analogous to the mutual inhibition explicitly implemented in each pool within an IAC network (see Fig. 6). Once a subset of the localist nodes in the IAC network are activated, this activity spread via the positive connections to nodes in the hub (which are also competing via inhibition) and from there to other localist nodes.

The network will eventually relax or settle into a locally optimal state satisfying as many relations as possible, with higher priority given to the stronger relations. Such a process is called "relaxing to a solution" or "relaxing to an interpretation," and always moves from a state satisfying fewer relations to a state satisfying more relations [59], or remains in a stable state with bias toward satisfying the input conditions. While IAC networks do not capture the learning aspects of schemata, the behavior of this structure does seem to capture the way our learned knowledge manifests in behavioral dispositions. Thus such structures provide an insight into a possible dynamics that is able to account for many aspects of cognition [32], [34], [41].

In comparing the two networks, the ERA unit implements exactly the same structure and dynamics as the IAC model, though with a relaxed requirement for the localist interpretability of each SOM node. Furthermore, the structure of the ERA unit is learned from the networks experience and so entities are grounded in whatever data stream is input to the network. By exhibiting the same context sensitivity, default assignment, and generalization properties as an IAC network, we can now see that the sensorimotor predictions of the ERA unit will equally be sensitive to context.

2) The Importance of Abstraction: Having considered the dynamics of the individual ERA units, we must now consider the importance of abstraction in sensorimotor prediction. From the perspective of object recognition, we can see that each layer of the hierarchy, by classifying combinations of the classifications of previous layers, is able to capture increasingly abstract entities. These abstractions need not all be directly related to the recognition of objects but will, in higher levels of the hierarchy, typically respond to things such as visual flow field directions, looming, and other sensorimotor effects. In combination with context sensitivity, such abstraction enables the prediction of the sensory consequences of actions with abstract general effects rather than simply relating to local object directed transformations.

B. Transparency and Conceptual Spaces

Unlike many evolved dynamical systems, the transparency of ERA allows an understanding of the behavior of the architecture

in terms of concepts. As already discussed, there is similarity between the classifications of SOMs in terms of their weight space and the geometric representations of CSs, ERA goes further by also reproducing the schemata-like priming of IAC networks thereby, demonstrating the use of these concepts. Of course, the interpretability of the SOMs as conceptual will be dependant on the interpretability of the input streams driving those SOMs. For example, where red, green, and blue pixel values form the input to a SOM, it is easily interpreted as a color space. As we move up the hierarchy, increasing levels of abstraction from the raw input, some work will be needed to analyze what the concepts are that play a role in the overt behavior of the system as a whole. Nevertheless, we can still understand the behavior of an ERA architecture even without explicitly tagging the concepts it uses.

C. Behavior and Sensory–Sensory Relations

So far the use of ERA to generate overt behavior in a robot has not been discussed, however, the use of priming in motor regions can easily be used to directly influence behavior. If the motor regions map something akin to the body-space used in the "modi" example, then motors can be activated to explicitly achieve those body postures. This would of course require an additional subsystem, and in some cases, such as the use of ERA to simulate potential actions and their consequences this may not be desirable. Decisions on how to direct overt behavior will be both dependant on the tasks it is being used to model and on the specific motor system and representation used. In some cases, it may also be beneficial to provide copies of the actual motor signals rather than a body-space thereby separating proprioceptive and motoric modalities.

ERA also goes beyond pure sensorimotor relations interpreted as the prediction of sensory input following motor actions, by allowing for sensory–sensory predictions. As an example, on seeing part of a car (say partially occluded) the system should be able to predict the sensory activity of the whole car including the occluded part following the spread of activation between the previously associated subparts of the car.

D. Development and Habituation

As an architecture that continuously learns, ERA models also display a developmental trajectory whereby new experiences scaffold new behavioral abilities. However, it is necessary to provide some initial behaviors to kick-start the exploration required in order to generate experience from which to learn. Such initial behavior could take the form of random exploration (perhaps implemented by adding weak noise to the motor regions), innate reactive responses, or could be guided by human–robot social interaction through various methods. One simple example would be to add an attention mechanism causing the robot to look at moving or changing objects in the visual field (as was used in the "modi" experiment).

Goal directed behavior is another aspect of the system that must be carefully designed in relation to what you want your ERA based models to do. One possibility is to include a reward modality of input to the architecture and block the motor responses associated with poor rewards, or use it to bias the dynamics of the ERA model toward rewarding states. This aspect of modeling with the ERA architecture has not been significantly explored yet and further work is required to explore and find satisfactory methods of achieving goal directed behavior. Care must be taken to provide appropriate initial experience for any ERA model as habit formation can quickly produce self-reinforcing repetitive behaviors. Nevertheless, habit formation is a property of human cognition, and is therefore not seen as a drawback of ERA-based models.

VII. Extensions to ERA

The ERA architecture as described is intended to form a set of modeling guidelines and methodologies for modeling in epigenetic robotics, and though we provide a formal implementation, we do not wish to overly constrain models. As such we anticipate many variations and extensions following the basic principles of operation that we have outlined. As an example of one potential variation and extension, we now summarize the use of reservoir systems as input filters and their relation to the biology of the cortex.

The cerebral cortex is evident early in embryonic development (from the five-vesicle stage) from which point its sheet-like growth in mammals is tremendous, leading to gyri, separated by sulci (folds caused by the skull restricting growth). While the extent of this folding varies in different species, regionalization of the cerebral cortex somewhat based upon these gyri is generally agreed upon. In adult human brains, different regions of cortex are associated with different functions though as Karmiloff-Smith *et al.* [20], [60], [61] has shown with extensive functional magnetic resonance imaging (fMRI) work, many functions such as language are globally processed in young children and gradually become locally processed by adulthood. For other regions, function is determined by connection to particular sensory modalities, for example the visual cortex is present very early and develops visual functionality quickly during normal development. However, as Sharma *et al.* [21] have shown, cutting the optic and auditory nerves and crossing them over in infant ferrets leads not only to relatively normal sight and hearing, but also the development of structures only ever normally present in the visual cortex, developing in the auditory cortex instead. ERA models would also display similar plasticity as each modality is processed in much the same way, but the resulting cognitive structure depends entirely on the associations based upon relationships present in the input streams. While this example is rather extreme, evidence of the plasticity of the cortex and its ability to reorganize both functionally and structurally following changes in input are persistent throughout the neuroscience literature. The general conclusion is that the brain is not like a Swiss army knife with functional modules genetically prespecified to emerge at certain stages of development, but rather that its structure is developmentally contingent upon its interactions with the world [15], [16].

Focusing at a different scale, and again following Swanson [53], the neocortex consists of the same number of layers throughout, six layers in both humans and rats while phylogenetically older parts of the cerebral cortex, such as the hippocampus only have three layers. In rats, as in humans, the

first (outer) layer of the neocortex consists mainly of wiring and has relatively few cell bodies, layer 2 and 3 typically contain small pyramidal neurons which project to other cortical regions in the same and different hemispheres respectively. Layer 4 consists mainly of granule cells which form local circuits, while layers 5 and 6 contain larger pyramidal neurons typically projecting descending connections to the brainstem, thalamus, and spinal cord, as well as to the motor system broadly defined. The precise makeup of these layers in terms of the density of cell bodies in each layer varies considerably in different regions of cortex. Projections to the thalamus from the cerebral cortex are reciprocal (thalamocortical loops) and topographically arranged. Other topographic loops exist between much of the cortex and other brain areas such as the basal ganglion which is hypothesized to be a centre for action selection [50], [62].

Following Mountcastle [63], the layered cortex is also vertically differentiated into cortical microcolumns, each consisting of between 10 and 100 000 cells. Microcolumns in rat somatosensory cortex typically consist of around 100 neurons [64], [65]. Lateral connectivity between columns is typically inhibitory and local while excitatory connectivity via layer 1 or via connectivity below layer 6 is typically between columns in different regions of cortex, or descending projections to subcortical areas of the brain and brain stem. For many, the cortical microcolumn is viewed as the basic computational unit of the cortex and accordingly provides the basis for our cortical model [48], [53], [63].

A. Abstract modeling of the Cortical Microcolumn

One of the major limitations of the ERA architecture is its poor ability to capture temporal and nonlinear relationships, which can be crucial to accurate prediction and behavior production. To address this shortcoming, and in relation to the underlying biology of the neocortex, dynamic reservoirs can be used as input-filters to some or even all of the SOMs in an ERA architecture.

Claiming that dynamic reservoirs are models of cortical microcolumns generally causes some concern to neuroscientists. For this reason, we must make it absolutely explicit that our aim here is not to mimic the specific circuitry and make up of cortical microcolumns, but rather to abstractly capture the following properties:

- cortical microcolumns are nonchaotic;
- cortical microcolumns do not display stable attractor dynamics (their activity quickly decays on cessation of input);
- input size to cortical microcolumns is sparse relative to the size of the microcolumn;
- the state space achieved by an active "firing" microcolumn is large and sensitive to its input [64], [65].

These properties of biological cortical microcolumns have very useful computational implications; first, by making highly nonlinear features linearly separable (much as a kernel warping function does in a support vector machine), and second, by acting as a fading memory [66]. These properties and the computational advantages they imbue are well documented in the reservoir computation literature for both the liquid state machines and echo state networks (ESN) [67]. As a starting point then, we use the ESN as a simple and very abstract model of a cortical microcolumn. The ESN we use is a discrete time neural network derived from a random weights matrix typically populated with 30% connectivity and adjusted so as to have a spectral radius < 1, i.e., $|\lambda_{\max}| < 1$, where $\lambda_{\max}$ is the eigenvalue of $\boldsymbol{w}$ which has the largest absolute value, this typically ensures that the resulting neural network implements a single null point attractor and so its dynamics are always input driven. The reservoir is then cycled according to the following standard equations.

The net input activity of discrete ESN

$$a_j = \sum y_i w_{ij} + i_i. \tag{5}$$

The output activity of a discrete ESN

$$y_j = \tanh(a_j). \tag{6}$$

All inputs to the ESN use the same update rules via a similar random weights matrix generated with 30% connectivity. As the output of a cortical microcolumn is hypothesized to be a classification of its input [48], yet a readerless ESN has no output. In using ESN's as input filters, their entire state should then be passed as input to the connected SOM (which should be significantly larger than the ESN), thus the SOM classifications are no longer interpretable as a conceptual space, classifying rather the region of state space in which the input filter ESN is currently in. Despite the lack of easy interpretation in terms of concepts, the same dynamics will follow as is found in the nonextended ERA architecture and so by analogy the resulting behavior of the system will remain predictable and transparent.

One of the problems with a reservoir systems approach is that because networks are randomly initialized, one can never know in advance if a particular random instantiation will be good as solving the problems you are interested in. Both Jaeger and Maass suggest that if you find a reservoir system unable to solve the problem you are interested in then you can either make larger reservoir systems (containing a greater range of dynamics hopefully more suited to your task), or alternatively you could evolve your reservoir in order to find one that is suited to your task [67], [68]. These suggestions are, however, unsatisfactory for two different reasons; first, evolving the reservoir would go against the idea of having a general purpose structure able to help solve (by linearising) a wide range of problems. Second, while making larger and larger reservoirs clearly does work, it would seem to be biologically implausible beyond some range of parameters. If reservoir dynamics are present in cortical microcolumns, as we and others suggest, then the size of a reservoir should not exceed the size of a cortical microcolumn. It should be noted here that the size of cortical microcolumns (in terms of the number of neurons) does indeed vary considerably in different regions of the mammalian brain with some microcolumns containing in excess of 1000 neurons [53]. So we can still have quite large microcircuits, but there is clearly an upper limit implied here. An alternative solution to the problem of finding a reservoir that produces the right kind of dynamics for your particular problem, whatever that problem may be is to use a network of reservoirs predicting each other's activity, and this is precisely what we think the cortex is doing.

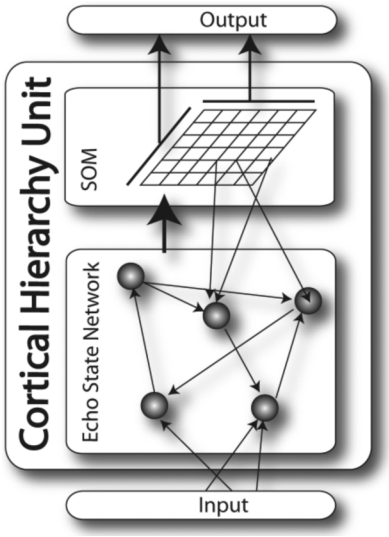

Fig. 7. Abstract model of a cortical microcircuit. Input perturbs an ESN reservoir, which is then read by a SOM. The SOM also provides an input to the ESN and the location of the winning SOM unit in SOM space is provided as the output of the unit.

Our solution to this problem rests on the use of feedback into a reservoir, achieved by the addition of echo state networks to the SOMs already present in ERA as shown in Fig. 7. As Jaeger and Maass have both demonstrated, feeding back the response of a trained readout as an input to reservoir typically enhances the ability of the resulting network to accurately perform the readout, and further allows for a reciprocal relationship between the inputs and readouts of that reservoir. For example, Jaeger [67] demonstrated that a readout trained to produce a value consistent with the frequency of an input, where feedback (both from reservoir to input, and from readout to reservoir) was used, was also able to produce a sign wave of the appropriate frequency in the input by clamping the value of the readout. In Morse *et al.* [69], we provide a detailed analysis of the effect of an external feedback signal and show that inputs that generally correlate with some (presumably nonlinear) feature of the input enhance the networks ability to detect such input features. We have also shown that such feedback reduces the detection of nonprimed features and so results in sustained inattentional blindness being displayed by the model. Used within an ERA architecture, the top–down feedback provides precisely this signal, focusing the ESN input filter appropriately to detect anticipated features of whatever input stream is driving the reservoir.

VIII. Conclusion

The ERA architecture as presented here forms a set of guidelines for the integration of SOM based modeling efforts into a system both capable of exhibiting a wide range of psychological effects and, at least in its nonextended version, one that operates transparently with concepts. Such transparency makes it relatively easy to extend the architecture with the integration of other systems beyond the current scope of ERA. What we believe we have achieved is an approach to modeling that can scale up beyond simple scenarios, is general in that it is not tailored to specific domains and tasks, and displays an ongoing developmental trajectory. Clearly, much of this remains to be demonstrated in future work and, equally clearly, there are limitations in the extent to which ERA satisfies each of these goals. Nevertheless, such an approach to modeling not only makes connections to a wide range of theories in the constructivist and sensorimotor paradigms, but demonstrates a simple method of the integration of many psychological phenomena, including development, within a single model. The basic architecture and variations thereof have already be used to demonstrate various phenomena which we have not discussed herein, examples include conditioned learning [70], various forms of priming, the relation between movement and orientation selectivity [71], sustained inattentional blindness [69], as well as the "modi" example discussed herein [42]. A great deal more work is required to further develop the ERA architecture and establish it as a modeling methodology. We therefore invite those interested to join us in our efforts to use and develop ERA further.

References

[1] A. Clark, *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*. London, U.K.: Oxford Univ. Press, 2008.
[2] K. L. Downing, "Neuroscientific implications for situated and embodied artificial intelligence," *Connect. Sci.*, vol. 19, no. 1, pp. 75–104, 2007.
[3] V. Gallese and G. Lakoff, "The brain's concepts: The role of the sensory-motor system in reason and language," *Cogn. Neuropsychol.*, vol. 22, pp. 455–479, 2005.
[4] A. Newell, "You can't play 20 questions with nature and win: Projective comments on the papers of this symposium," in *Vis. Inform. Process.*. New York: Academic, 1973, pp. 135–183.
[5] A. Noë, *Action in Perception*. Cambridge, Mass: MIT Press, 2004.
[6] S. Harnad, *Symbol Grounding Problem Physica D*, vol. 42, no. 1–3, pp. 335–346, 1990.
[7] T. Ziemke, "Rethinking grounding," *Understand. Rep. Cogn. Sci.*, pp. 177–190, 1999.
[8] R. A. Brooks, "A robust layered control system for a mobile robot," *IEEE J. Robot. Autom.*, vol. 2, no. 1, pp. 14–23, 1986.
[9] R. D. Beer, "Dynamical approaches to cognitive science," *Trends Cogn. Sci.*, vol. 4, no. 3, pp. 91–99, 2000.
[10] T. Van Gelder, "The dynamical hypothesis in cognitive science," *Behav. Brain Sci.*, vol. 21, no. 5, pp. 615–628, 1998.
[11] J. C. Bongard, "Incremental Approaches to the Combined Evolution of a Robot's Body and Brain," Ph.D. thesis, Faculty of Mathematics and Science, Univ. Zurich, Zurich, Switzerland, 2003.
[12] R. Pfeifer, J. Bongard, and S. Grand, *How the Body Shapes the Way We Think: A New View of Intelligence*. Cambridge, MA: MIT Press, 2007.
[13] S. Collins, A. Ruina, R. Tedrake, and M. Wisse, "Efficient bipedal robots based on passive-dynamic walkers," *Science*, vol. 307, no. 5712, p. 1082, 2005.
[14] A. Clark and D. J. Chalmers, "The extended mind," *Analysis*, vol. 58, pp. 10–23, 1998.
[15] S. Oyama, *Evolution's Eye: A Systems View of the Biology-Culture Divide*. Durham, NC: Duke Univ. Press, 2000.
[16] S. Oyama, *The Ontogeny of Information: Developmental Systems and Evolution*. Durham, NC: Duke Univ. Press, 2000.
[17] S. R. Harnad, *Categorical Perception: The Groundwork of Cognition*. Cambridge, U.K.: Cambridge Univ Press, 1990.
[18] A. M. Glenberg and M. P. Kaschak, "Grounding language in action," *Psychonomic Bulletin Rev.*, vol. 9, no. 3, pp. 558–565, 2002.
[19] J. Parthemore and A. F. Morse, "Reclaiming symbols: An enactive account of the inter-dependence of concepts and experience," *Pragmatic. Cogn.*, vol. 18, no. 1, 2010.
[20] A. Karmiloff-Smith, "Why Babies' Brains Are Not Swiss Army Knives," in *Alas, Poor Darwin*, H. Rose and S. Rose, Eds. London, U.K.: Jonathan Cape, 2000, pp. 144–156.
[21] J. Sharma, A. Angelucci, and M. Sur, "Induction of visual orientation modules in auditory cortex," *Nature*, vol. 404, pp. 841–847, 2000.
[22] A. Noë, *Out of Our Heads*. New York: Hill & Wang, 2009.
[23] K. O'Regan and A. Noë, "A sensorimotor account of visual perception and consciousness," *Behav. Brain Sci.*, vol. 24, pp. 939–1011, 2001.

[24] J. J. Gibson, *The Ecological Approach to Visual Perception*. Boston, MA: Houghton Mifflin, 1979.
[25] L. W. Barsalou, P. M. Niedenthal, A. K. Barbey, and J. A. Ruppert, "Social embodiment," *Psychol. Learn. Motivation: Adv. Res. Theory*, vol. 43, pp. 43–92, 2003.
[26] F. Strack, L. L. Martin, and S. Stepper, "Inhibiting and facilitating conditions of the human smile: A nonobtrusive test of the facial feedback hypothesis," *J. Personality Social Psychol.*, vol. 54, no. 5, pp. 768–777, 1988.
[27] M. Chen and J. A. Bargh, "Consequences of automatic evaluation: Immediate behavioral predispositions to approach or avoid the stimulus," *Personality Social Psychol. Bulletin*, vol. 25, no. 2, p. 215, 1999.
[28] L. B. Smith and L. Samuelson, "Objects in space and mind: From reaching to words," in *Thinking Through Space: Spatial Foundations of Language and Cognition*, K. Mix, L. B. Smith, and M. Gasser, Eds. London, U.K.: Oxford Univ. Press, 2010.
[29] P. Gärdenfors, *Conceptual Spaces: The Geometry of Thought*. Cambridge, MA: MIT Press, 2004.
[30] E. Rosch, "On the internal structure of perceptual and semantic categories," *Cogn. Develop. Acquisition Lang.*, vol. 12, p. 308, 1973.
[31] T. Kohonen, "The self-organizing map," *Neurocomputing*, vol. 21, no. 1–3, pp. 1–6, 1998.
[32] V. Bruce, A. M. Burton, and I. Craw, "Modelling face recognition," *Philosoph. Trans. Roy. Soc., B, 335, 121*, vol. 128, 1992.
[33] A. M. Burton, "Learning new faces in an interactive activation and competition model," *Vis. Cogn.*, vol. 1, no. 2, pp. 313–348, 1994.
[34] A. M. Burton, V. Bruce, and P. J. B. Hancock, "From pixels to people: A model of familiar face recognition," *Cogn. Sci.*, vol. 23, no. 1, pp. 1–31, 1999.
[35] G. Westerman and E. R. Miranda, "Modelling the development of mirror neurons for auditory-motor integration," *J. New Music Res.*, vol. 31, pp. 367–375, 2003.
[36] P. Li, I. Farkas, and B. MacWhinney, "Early lexical development in a self-organizing neural network," *Neural Netw.*, vol. 17, no. 8–9, pp. 1345–1362, 2004.
[37] P. Li, X. Zhao, and B. MacWhinney, "Dynamic self-organization and early lexical development in children," *Cogn. Sci.*, vol. 31, no. 4, pp. 581–612, 2007.
[38] J. Mayor and K. Plunkett, "A neurocomputational account of taxanomic responding and fast mapping in early word learning," *Psychol. Rev.*, vol. 117, no. 1, pp. 1–31, 2010.
[39] D. Caligiore, A. Borghi, D. Parisi, and G. Baldassarre, "TRoPICALS: An embodied newural-network model of experiments on compatibility effects," *Psychol. Rev.*, 2010.
[40] A. Clark, *Microcognition: Philosophy, Cognitive Science, and Parallel Distributed Processing*. Cambridge, MA: MIT Press, 1991.
[41] M. Page, "Connectionist modelling in psychology: A localist manifesto," *Behav. Brain Sci.*, vol. 23, no. 04, pp. 443–467, 2001.
[42] A. F. Morse, T. Belpaeme, A. Cangelosi, and L. B. Smith, "Thinking with your body: Modelling spatial biases in categorization using a real humanoid robot," in *32nd Annu. Conf. Cogn. Sci. Soc.*, Portland, OR, 2010.
[43] J. Piaget, *The Origins of Intelligence in Children*. New York: Norton, 1963.
[44] D. A. Baldwin, "Early referential understanding: Infant's ability to recognize referential acts for what they are," *Develop. Psychol.*, vol. 29, pp. 832–843, 1993.
[45] G. Metta, G. Sandini, D. Vernon, L. Natale, and F. Nori, "The iCub humanoid robot: An open platform for research in embodied cognition," in *Proc. IEEE Workshop Perform. Metrics Intell. Syst.*, Washington, DC, 2008.
[46] V. Tikhanoff, A. Cangelosi, P. Fitzpatrick, G. Metta, L. Natale, and F. Nori, "An open-source simulator for cognitive robotics research: The prototype of the icub humanoid robot simulator," in *Proc. IEEE Workshop Perform. Metrics Intell. Syst.*, Washington, D. C., 2008.
[47] M. J. Frank, L. C. Seeberger, and R. C. O'Reilly, "By carrot or by stick: Cognitive reinforcement learning in Parkinsonism," *Science*, vol. 306, no. 5703, p. 1940, 2004.
[48] J. Hawkins and S. Blakeslee, *On Intelligence*. New York: Times Books, 2004.
[49] J. Tani, "On the interactions between top-down anticipation and bottom-up regression," *Frontiers Neurorobot.*, vol. 1, 2007.
[50] T. J. Prescott, "Forced moves or good tricks in design space? Landmarks in the evolution of neural mechanisms for action selection," *Adapt. Behav.*, vol. 15, pp. 9–31, 2007.
[51] E. G. Jones and T. P. S. Powell, "An anatomical study of converging sensory pathways within the cerebral cortex of the monkey," *J. Anatomy*, vol. 93, pp. 793–820, 1970.
[52] M. W. Brown and J. P. Aggleton, "Recognition memory: What are the roles of the perirhinal cortex and hippocampus?," *Nature Rev. Neurosci.*, vol. 2, pp. 51–61, 2001.
[53] L. W. Swanson, *Brain Arch.Understanding the Basic Plan*. London, U.K.: Oxford Univ. Press, 2003.
[54] K. L. Downing, "Predictive models in the brain," *Connect. Sci.*, 2008.
[55] J. L. McClelland and D. E. Rumelhart, "An interactive activation model of context effects in letter perception: Part 1. An account of basic findings," *Psychol. Rev.*, vol. 88, p. 375407, 1981.
[56] J. L. McClelland, D. E. Rumelhart, and T. P. R. Group, *Parallel Distributed Processing: Psychological and Biological Models*. Cambridge, MA: MIT Press, 1986.
[57] S. Fidler, G. Berginc, and A. Leonardis, "Hierarchical statistical learning of generic parts of object structure," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, New York, 2006.
[58] S. Fidler and A. Leonardis, "Towards scalable representations of object categories: Learning a hierarchy of parts," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, Minneapolis, MN, 2007.
[59] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," *Proc. Nat. Acad. Sci.*, vol. 79, no. 8, p. 2554, 1982.
[60] A. Karmiloff-Smith, "Development itself is the key to understanding developmental disorders," *Trends Cogn. Sci.*, vol. 2, no. 10, pp. 389–398, 1998.
[61] A. Karmiloff-Smith, J. H. Brown, S. Grice, and S. Paterson, "Dethroning the myth: Cognitive dissociations and innate modularity in Williams syndrome," *Develop. Neuropsychol.*, vol. 23, no. 1&2, pp. 227–242, 2003.
[62] M. D. Humphries, R. D. Stewart, and K. N. Gurney, "A physiologically plausible model of action selection and oscillatory activity in the basal ganglia," *J. Neurosci.*, vol. 26, no. 50, p. 12921, 2006.
[63] V. B. Mountcastle, "An Organizing Principle for Cerebral Function: The Unit Model and the Distributed System," in *The Mindful Brain*, Edelman and Mountcastle, Eds. Cambridge, MA: MIT Press, 1978.
[64] A. Gupta, G. Silberber, M. Toledo-Rodriguez, C. Z. Wu, Y. Wang, and H. Markram, "Organizing principles of neocortical microcircuits," *Cellular Molecular Life Sci.*, 2002.
[65] H. Markram, Y. Wang, and M. Tsodyks, "Differential signaling via the same axon of neocortical pyramidal neurons," *Nat. Acad. Sci.*, pp. 5323–5328, 1998.
[66] W. Maass, T. Natschlager, and H. Markram, "Real-time computing without stable states: A new framework for neural computation based on perturbations," *Neural Comput.*, vol. 14, no. 11, pp. 2531–2560, 2002.
[67] H. Jaeger, "Tutorial on training recurrent neural networks, covering BPTT, RTRL, EKF and the 'echo state network' approach," GMD-Forschungszentrum Informationstechnik 2002.
[68] W. Maass, T. Natschlager, and H. Markram, "A Model for Real-Time Computation in Generic Neural Microcircuits," in *Advances in Neural Information Processing Systems 15*. Cambridge, MA: MIT Press, 2003, pp. 213–220.
[69] A. F. Morse, R. Lowe, and T. Ziemke, "Manipulating space: Modelling the role of transient dynamics in inattentional blindness," *Connect. Sci.*, vol. 21, no. 4, pp. 275–295, 2009.
[70] A. F. Morse and M. Aktius, "Dynamic liquid association: Complex learning without implausible guidance," *Neural Netw.*, vol. 22, no. 1, pp. 875–889, 2009.
[71] A. F. Morse and T. Ziemke, "Action, detection, and perception: A computational model of the relation between movement and orientation selectivity in the cerebral cortex," in *Proc. CogSci 2009—31st Annu. Conf. Cogn. Sci. Soc.*, Amsterdam, The Netherlands, 2009.

Anthony F. Morse received the D.Phil. degree from the University of Sussex, Sussex, U.K., in 2006 before working as a Postdoctoral Researcher for three years at Skovde University, Skovde, Sweden.

He is currently a Senior Research Fellow at the University of Plymouth, Plymouth, U.K., and a member of the Centre for Robotics and Neural Systems. He works on the EU FP7 project "ITALK: Integration and Transfer of Action and Language Knowledge in Robotics," a multipartner research project on action and language learning on the humanoid robot iCub. His research interests include cognitive systems, sensorimotor perception, developmental psychology, and the challenge of putting it all together.

Joachim de Greeff received the M.Sc. degree from Utrecht University, Utrecht, The Netherlands, in 2007, with a thesis on the evolution of communication in mobile robots. He is currently working towards the Ph.D. degree in modeling concepts in artificial systems at the University of Plymouth, Plymouth, U.K.

He spent some time working at the Laboratory of Autonomous Robotics and Artificial Life in Rome and is currently working at Plymouth University. His research interests include cognitive systems, conceptual learning, evolutionary robotics, and general questions regarding AI.

Tony Belpaeme received the M.Sc. degree in electronic engineering from, KHBO, Belgium, in 1994. He received the Ph.D. degree from the Vrije Universiteit Brussel (VUB), Amsterdam, The Netherlands.

He is currently an Associate Professor (Reader) in Intelligent Systems at the University of Plymouth, Plymouth, U.K., where he is a member of the Centre for Robotics and Neural Systems. Before joining Plymouth in 2005, he worked as a Postdoctoral Researcher and Lecturer at the VUB. His research interests include cognitive systems, human–robot interaction, conceptualization, and the interaction between language and cognition.

Angelo Cangelosi received the Ph.D. degree in psychology and computational modeling from the University of Genoa, Genoa, Italy, in 1997, while also working as a visiting scholar at the National Research Council, Rome, the University of California San Diego, La Jolla, CA, and the University of Southampton, Southampton, U.K.

He is currently a Professor of Artificial Intelligence and Cognition at the University of Plymouth, Plymouth, U.K., where he leads the Centre for Robotics and Neural Systems. He has produced more than 140 scientific publications and has been awarded numerous research grants from the United Kingdom and international funding agencies.