How to design emergent models of cognition for application-driven artificial agents

Serge Thill and David Vernon

Interaction Lab,
School of Informatics,
University of Skövde,
541 28 Skövde, Sweden
E-mail: {serge.thill; david.vernon}@his.se

Emergent models of cognition are attractive for artificial cognitive agents because they overcome the brittleness of systems that are fully specified in axiomatic terms at design time, increasing, for example, the ability to deal with uncertainty and unforeseen events. When the agent is created to fulfil specific requirements defined by a given application, there is an apparent conflict between the emergent (*i.e.* self-defining) nature of the agent's behaviour and the pre-specified (*i.e.* axiomatically-defined) nature of the requirements.

Here, we develop a framework for the design of emergent models of cognition whose behaviour can be shaped to fulfil application requirements while retaining the desired characteristics of emergence. We achieve this by viewing the artificial agent as forming an eco-system with the environment in which it is deployed. Consequently, the objective function that determines the agent's behaviour is cast in terms that factor in interaction with the environment (while not being controlled by it) and therefore implicitly includes the application requirements.

This framework is particularly relevant to application driven research where artificial agents are designed to interact with humans in a certain manner. We illustrate this with the example of robot-enhanced therapy for children with autism spectrum disorder.

Keywords: Emergent models of cognition, application-driven research, robot-enhanced therapy

1. Models in cognitive science and new challenges from embodied theories

Typically, models of cognition serve the study of cognition itself. McClelland put it rather succinctly when he argued[1] that we should think of models as "tools for exploring the implications of ideas" (p 12), even though they are necessarily simplifications of the real thing. Traditional cognitive science is replete with examples from many paradigms (whether symbolic, subsymbolic, Bayesian, or others) that have successfully done just that.

Models have been useful in the cognitive sciences because they tend to be *explanatory* models of cognitive mechanisms. With the advent of embodied theories of cognition (according to which human cognition cannot simply be entirely reduced to abstract, amodal symbol manipulation), however, new challenges have appeared. These theories often lead, for instance, to a need for "embodying" the model. This is typically done with an artificial agent, which can be simulated or real.

These approaches have significant hurdles to overcome if the aim is to provide an explanatory model of human cognition. Robotic models, for instance, are limited by the robotic hardware available, in particular the fact that these do not provide anything resembling a human body, even if the robots used are described as "humanoid" and "embodied". Models "embodied" in simulated environments, meanwhile, lose a significant degree of realism since simulators simply cannot approximate the complex physics of reality. Like robotic models, they raise the question of whether a (simulated) non-human body is an acceptable abstraction of a human body but extend the issue to the environment itself.

It has been argued repeatedly[2,3] that an advantage of robotic models of cognition is that they are forcibly integrated, since every process from sensory perception to the mechanisms of the cognitive behaviours of interest needs to be modelled. This is rarely true, however, as readily illustrated by – for instance – the complexities of computer vision forcing modellers to take short cuts in obtaining visual inputs, for example by assuming a process which returns coordinates of objects of interest[4] or by using coloured, easily-discriminable objects[5]. The promised "complete" modelling, from sensory perception to higher-level cognition is thus not necessarily given and adds to the non-human embodiment used in simplified environments (even when real robots are used) to create a model embodied in something that is entirely different from a human experience. In particular, although sensorimotor aspects are thought to fundamentally shape higher cognition, they are often the first to be simplified.

Additionally, it is not a given (even though this is often assumed) that merely instantiating a model in a robotic body overcomes the limitations of traditional symbolic approaches to (strong) AI that such models typically intend to address (by virtue of taking an embodied, as opposed to amodal symbol-processing, view of human cognition). For instance, although they are often presented as overcoming the problems illustrated by the Chinese Room argument[6], they tend to ignore that Searle, in the original paper[7], already rejected the "robot reply" – collecting inputs from sensors and manipulating them to produce motor outputs – as a way of achieving an AI to which genuine understanding and mental states can be attributed (see 6 for a fuller discussion).

However, this is not to say that they have no explanatory power (or indeed no utility!) at all. For example, even strongly abstracted models of sensorimotor mechanisms can provide insights into minimal requirements for the cognitive process of interest[3]. While not necessarily creating an account of cognitive mechanisms *per se*, such strategies can nonetheless constrain the search space. Similarly, theoretical models can explore how *changes* in embodiment might affect cognitive processes that depend on it[8].

Overall, therefore, the realisation that human cognition is embodied to a degree that cannot be abstracted way makes "exploring the implications of ideas" with computational models much more challenging. At the same time, however, robots also create an entirely new *raison d'être* for cognitive models, which does not rely on explanatory insights about human cognition. The remainder of this paper is dedicated to the discussion of such an example.

2. Non-explanatory uses of models

Issues in explanatory power notwithstanding, technological advances in a number of areas dealing with human-machine interaction[9–11] give models of cognitive mechanisms an important reason of existence: to interact proficiently with humans, such machines need at least a rudimentary *Theory of Mind* (ToM); an internal model that can be used to estimate mental states of humans, in particular their intentions, expectations and predicted reactions to actions by the agent[10,12].

Additionally, to create machines with a given human ability necessitates a mechanistic model of this ability. Whether or not the model provides an adequate explanation of the mechanisms underlying the cognitive behaviour in a human is irrelevant here; it is sufficient that the postulated mechanisms can be exploited by the artificial agent. In contrast with McClelland's take on the utility of models, such models (although of human cognitive phenomena) do not need to possess any explanatory power; appropriate behaviour according to specification when instantiated in an artificial agent, no matter how biologically implausible the underlying mechanisms, is sufficient.

A particular human phenomenon of interest here is the *emergence* of appropriate cognitive behaviours. It is generally not possible to fully specify the behaviour that artificial cognitive agents ought to display since that would require complete knowledge of every situation they are likely to encounter. Emergent models therefore do not specify agent behaviour axiomatically at design time; rather the agent discovers appropriate behaviour[a] itself as a partial consequence of its interaction with the world in which it is embedded, brought about by the system's self-organizing system dynamics. This reduces the brittleness of its behaviour while increasing its ability to adapt to new situations or deal with events that the system designers had not foreseen.

3. Emergent models in purpose-built artificial agents

The prevalent current trend[b] in artificial cognitive systems research is to focus on the creation of artificial agents for specific purposes (often defined by given application scenarios). This is in contrast with previous, more explorative, research lines in which the agents are for instance used to illustrate principles of morphological computation[16] or embodiment.[17]

As an example, consider the use of (partially autonomous) cognitive humanoid robots in therapeutic interventions with children who have autism spectrum disorder (ASD)[18]: so-called robot-assisted therapy (RAT) and robot-enhanced therapy (RET). Although significant research problems (pertaining for instance to the controller of the robot[10]) remain, it is the end users (the therapists using these robots) who determine what the behaviour of the robot should be. These requirements drive research on, for instance, suitable robot control.

It is therefore interesting to reflect on the relevance of *emergent* models of cognition in a context in which the required behaviour is defined and driven by a specific application, and the primary mechanistic purpose of a model is to ensure these requirements are satisfied. After all, the application-driven specification of the cognitive behaviour of a robot (or

[a] This emergent behaviour derives from a process whereby the agent continually strives to make sense of it environment through its interaction with that environment. Consequently, the system's understanding of its world is inherently specific to the form of the system's embodiment and is dependent on the system's history of interactions, *i.e.*, its experiences. This process of making sense of its environmental interactions is one of the foundations of a branch of cognitive science called *enaction*[13,14] and is related to the concept of radical constructivism[15].

[b] This trend is evident, for example, in current EU Horizon2020 funding schemes, which place an increased importance on application-driven research; that is, research whose content is specified by the needs of an application.

other artificial agent) seems at first glance to be at odds with emergent behaviour that, by definition, is not fully defined a priori.

We argue, using the example case of RET for illustration, that there is a useful place for emergent models in an application-driven world despite the apparent contradiction in that the behaviour of the system has to be both emergent and specified a priori.

In RET, most research and interventions are focused either on basic mechanisms, such as imitation, or various basic cognitive-behavioural skills, such as joint attention and turn taking, both of which are related to the core symptoms (e.g. communication and social interaction), disabilities, and impairments (e.g., lack of social skills). Indeed, the severity of autism is correlated with impaired imitation skills, joint attention, and turn taking.[19] As such, children with autism fail to imitate and to have joint attention episodes from an early age; a salient diagnostic marker for the disorder.[20] Recently, researchers have found that for children with ASD, imitation and joint attention acquisition improves in settings in which technological tools are involved[21–26]. Taking into account that ASD patients tend to learn more from the interaction with technology rather than from the interaction with the human beings, robots might have the potential to be used in ASD therapies as intermediaries between human models and ASD patients[27].

Although strongly scripted, successful therapy depends on the robot adapting to the child's behaviour and engaging in a behaviour of its own that fosters continued interaction. That is, the robot has to anticipate the child's behaviour and act effectively in the ensuing interaction to fulfil the application-driven requirement of promoting robot-child engagement in the face of uncertainty in interaction. It must therefore be able to successfully predict outcomes of actions (thereby managing or reducing uncertainty) while maintaining interaction with the child. This is the essence of cognition and herein lies the key reason for the use of emergent cognitive systems in application-driven scenarios: even though the fact that all required behaviours are pre-defined by the application might suggest a pre-defined axiomatic model of cognition, it is not possible to specify the use-case scenarios to the extent such an approach would require because it is not possible to control or predict exactly how a child will interact with a robot at all times during the intervention. Emergent models however, allow behaviour to arise *without having to define it functionally in axiomatic terms.*

The catch is the following: An agent's emergent behaviour is not controlled directly but is modulated by an embedded (or intrinsic) value system[28], creating a challenge for the deployment of emergent systems in predefined applications. This value system provides the motives that determine the agent's actions and guide its development. In particular, it mediates the saliency of environmental stimuli, flagging the occurrence of important stimuli, and triggering the formation of goals[29]. The research problem is therefore to (a) identify a value system that satisfies the conditions of emergence and facilitates application-compatible agent-environment interaction, and (b) cast it in *quantitative* form so that it can be used to modulate the robot's emergent behaviour.

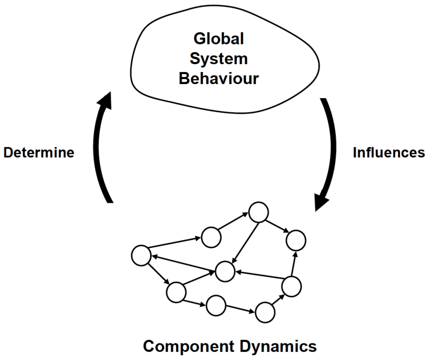

Fig. 1. Circular causality in cognition: global system behaviour influences the local components that determine the global behaviour in the first place (from 30).

To address it, we make two propositions. First, artificial agents with emergent cognitive abilities should be considered as forming an eco-system with the environment in which they are deployed, much in line with situated views of human cognition. Consequently, the value system should be cast in terms of the predictive interaction with this environment. Second, the dynamics that govern the emergent self-organization of the system should be cast as a quantitative objective function that has the value system as one of its parameters. This objective function can be viewed as one or more policies that determine how each components in the system changes. It does not necessarily have to be fixed but can be adapted using, *e.g.*, reinforcement learning. The value system then shapes the objective function that governs the self-organization of the agent's constituent components which, in turn, give rise to the global systems behaviour.

This is effectively an instance of *circular causality*[31] (see Fig. 1): a cognitive agent's global behaviour is shaped by the dynamics of its local constituent components; simultaneously, these dynamics are influenced by the global behaviour[c]. To achieve desired behaviour by an artificial agent, we *shape and manipulate these local components.*

There are two key points to notice here. One is that the value system operates at the level of the global system, modulating the agent's behaviour and interaction with its environment. The other is that the objective function, parameterized in part by the state of the value system, operates at the level of the component dynamics and thereby governs the activity of the local components. Collectively, these component dynamics determine the global system behaviour in a circularly causal loop.

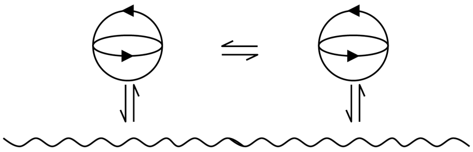

Fig. 2. Maturana and Varela's[34] ideogram depicting interaction between two agents, highlighting that interaction is a shared activity. (Picture from 30).

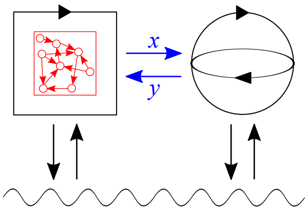

Fig. 3. Schematic of interaction between an artificial (left, square indicates that the agent is artificial) and another (right) agent that combines the concepts underlying Figs. 1 and 2. The appropriate behaviour emerges through appropriate configuration of the *local components* (in red), yet the behavioural requirements are defined in terms of the *interaction between the agents* (in blue).

That the application constrains the emergent behaviour is then helpful since it requires the researcher to be very explicit about the desired interactions (again, without having to define them functionally), thus reducing the

[c] Closely related concepts include continuous reciprocal causation[32] and recursive self-maintanence[33].

space in which the appropriate objective function can be found. Returning to the example of RET, a value system reflecting the need to maintain appropriate interactions with the child (allowing the system to increase its predictive capacity and tolerance to uncertainty) facilitates the emergence of interactive behaviours that also satisfy the application requirements, *i.e.* keeping the child engaged.

The nature of interactive behaviour in emergent systems is captured well in Maturana's and Varela's ideograms[34] (for a more complete discussion, see 30) of self-organising and self-maintaining enactive systems. When two such agents exist, they are coupled in terms of the *shared activity* that is their interaction (see Fig. 2). When building an artificial agent, the desired behaviour is typically *specified in terms of this interaction.* Here, we therefore primarily consider artificial agents built to interact with humans for two reasons: (1) artificial agents are seldom created to exist in complete isolation from humans; interaction is often part of their purpose, and (2) interacting with another human intrinsically introduces uncertainty and events that cannot be anticipated in advance. As argued before, these conditions make the use of emergent models particularly relevant in the first place.

We have now prepared the scene for emergent models of cognitive behaviour: *the dynamics of the local components must lead to global behaviour that satisfies the requirements*[d] *defined by the desired interactions, and the emergent system's value system should be designed to favour these interactions, modulating the local dynamics to promote this global behaviour* (Fig 3). We now build on this characterisation to describe a framework for emergent models of cognition in application-driven scenarios.

4. How to design emergent models of cognition in purpose-built artificial agents

First, we note that we can describe the behaviour of the artificial agent in the terms of dynamical systems:

$$\dot{x} = \Phi\left(x, p, \eta\right) \tag{1}$$

[d] A particular requirement could for instance be that this interaction be maintained; in other words, that it is undesirable for the artificial agent to either not produce any behaviour or to elicit no response from the interactant.

Here, x denotes system state, p represents the system parameters, and η is a noise term.[35]

We can also describe the interaction between two agents X (artificial) and Y (human, say) from the perspective of the artificial agent:

$$P = F_n\left(x\right) \approx \tilde{y} \tag{2}$$

$$M = B_n\left(\tilde{y}\right) \approx x \tag{3}$$

F_n is a family of n forward models[e] that describes the predicted perception P of the agent X given some state x. It effectively says what the agent expects to see — of itself and of agent Y — if it executes some motor program corresponding to the state x. Here, y represents the actual behaviour intended by agent Y, while $\tilde{y}$ denotes the aspects of that behaviour perceivable by the artificial agent. In other words, this model allows the agent to *predict the outcome of its own actions and the actions of agent Y in terms of the possible action states.* It will, in practice, be learned using any appropriate choice amongst many existing approaches. This prediction can be compared to $\tilde{y}$, what it is actually perceiving, i.e., the behaviour of agent Y.

B_n, then, is a family of models similar to inverse models that define what action M the agent should carry out, given the observation $\tilde{y}$. This can be compared to what the agent X is currently doing, i.e. its state x. These models therefore allow the agent to *verify the appropriateness of its own actions* - if its actual actions do not match with what it should have been doing, then the actions are not appropriate. Consequently, Eqn. 3 is therefore given (and formulated[f]) by the application-driven requirements.

The designer of a system can specify the desired behaviour of the artificial agent through a suitable family of B_n models. In the case of RET, for example, B_n would encapsulate the requirements of the end-users (the therapists) and describe the desired ways in which the robot should interact with the child (for instance, through joint attention, imitation, and turn taking). In 38, we present a detailed example of such an interaction specification.

[e] Note that this characterization of a forward model differs slightly from the conventional notion. For example, the HAMMER architecture uses inverse models to identify the required motor commands required to achieve some internally-generated goal and forward models to predict the perceptual outcome of executing those commands.[36,37] In this case, however, the goal state is not a perception of the behaviour of agent X, but of the interactant agent Y.

[f] at least to some functional degree; possible refinements can still be learned later.

We can now define an value system that the artificial agent seeks to maximise. Effectively, a successful agent will have developed behaviour that ensures it can both accurately predict the outcome of its own actions in terms of the interactant's behaviour *and* elicit appropriate behaviour from the interactant. Such a value system function Ψ can be formulated as follows.

$$\Psi\left(F_n\left(x\right), B_n\left(\tilde{y}\right)\right) \tag{4}$$

Given this value system function, we can now revisit the system behaviour as described by Eqn. 1. Specifically, replacing p with Ψ ensures that the behaviour of the system is now modulated by an objective function system based on the value system:

$$\dot{x} = \Phi\left(x, \Psi, \eta\right) \tag{5}$$

Since Ψ includes an evaluation of how well the agent's behaviour matches the task-specific requirements as specified by the B_n models, these requirements now shape what is otherwise emergent behaviour determined by the internal component dynamics of the artificial agent. This completes the core of what we have set out to do: Eqn. 5 effectively encapsulates a framework for the emergence of cognitive behaviour in an artificial agent that satisfies the requirements set by a given application. It also captures the essence of the circular causality described above, a key characteristic of emergent cognitive systems. The downward causality of the global system behaviour is captured by some function Ψ of (a) the agent's goal behaviour B as a function of its perceptions of the agent it is interacting with, and (b) its predicted perception of the behaviour of its interactant agent Y as a function of its possible behaviour x. The upward causality is the self-organizing function Φ which produces the component dynamics and hence the global behaviour, where this function has the downward value system state factored into it as a parameter.

Discovering a useful shape for Φ — necessary to ensure that the changes in the agent's behaviour bring it closer to the desired behaviour — is not trivial. We do however note that the problem might be usefully cast as a reinforcement learning problem[39]. The ability to maintain appropriate interactions with another agent in particular forms the basis of the reward signal which can guide modifications to Φ. Successfully maintaining such interactions, given a set of B_n models, would then also facilitate the emergence of interactive behaviours that satisfy the application requirements.

5. Discussion

There are (at least) three aspects worth considering to take this framework forward. The first is that Eqns. 2 and 3 are based on observable actions from both the artificial agent and the interactant. These actions are not instantaneous events but occur over some time – a fact that detailed implementations of the F_n and B_n models cannot ignore.

The second is that the framework at present purposefully ignores any implementation of a theory of mind (ToM) for the artificial agent.[10,12] This is because we wanted to highlight that the framework can be specified based solely on observable interactions and an approximation $\tilde{y}$ of what the underlying states of the interactant are (Eqn. 2). ToM models can be used to *predict* the states y, which in turn may be beneficial for the learning of a well-functioning family of forward models F_n.

The third concerns our replacing p with Ψ in Eqn. 1 – an act which requires more discussion in the context of autonomy. Emergent cognitive systems exhibit a considerable degree of autonomy. While there are many different types of autonomy[30], two are particularly relevent to the present discussion: behavioural and constitutive autonomy. Behavioural autonomy[g] focusses on the external characteristics of the system while constitutive autonomy focusses on the internal organization and the organizational processes that manage to keep the system viable and autonomous[42]. The constitutive-behavioural distinction is sometimes cast as a difference between *constitutive* processes and *interactive* processes[43,44]. Constitutive processes deal with the system itself, its organization, and its maintenance as a system through on-going processes of self-construction and self-repair. On the other hand, interactive processes deal with the interaction of the system with its environment. Both processes play complementary roles in autonomous operation of the system.

In this paper, we have focussed on the behavioural aspects of the agent, factoring in the value system into the objective function that governs the self-organization of the agent and hence its emergent behaviour. In doing so, we replaced p with Ψ in Eq. 1, and it can be argued that we were perhaps a little too quick to do so. The intrinsic value system Ψ modulates behaviour based on cognitive interaction with the environment; we might refer to it as a *behavioural value system.* However, there is different, *endogenous* value system that modulates the constitutive processes in the agent and

[g] Behavioural autonomy is sometimes referred to as adaptive or cognitive autonomy to reflect its focus on interactive processes rather than constitutive processes[40,41].

is equally important to the autonomy of the agent and its emergent self-organization. This value system function would have been responsible for adaptively setting the parameter values p in Eq. 1. So, to complete the picture, we must recognise that we need to include a second *constitutive* value system function, Ψ', to complement the *behavioural* value system function, Ψ, discussed above.

6. Conclusion

We have illustrated a symbiotic benefit between emergent systems and given application niches: the former are necessary to deal with the inherent uncertainty in deploying artificial systems to, for instance, interact with humans while the latter facilitate the clear specification of an objective function that in turn facilitates the emergence of desired behaviour.

The framework we have developed here capitalises on these insights to facilitate the use of emergent models of cognition in artificial agents created for specific applications. In the larger context of models of human cognition, it is an example that builds on properties of human cognition – emergence and autonomy – and has a clear use for ToM mechanisms. This highlights that the utility of models of human cognition goes beyond the exploration of the consequences of ideas (although this of course remains another important use of these models); they are also a necessity for the design of artificial agents that interact with humans.

7. Acknowledgements

This work has been supported by the EC FP7-ICT project DREAM (www.dream2020.eu), project number 611391.

References

1. J. L. McClelland, The place of modeling in cognitive science, *Topics in Cognitive Science* **1**, 11 (2009).
2. A. F. Morse, C. Herrera, R. Clowes, A. Montebelli and T. Ziemke, The role of robotic modelling in cognitive science, *New Ideas in Psychology* **29**, 312 (2011).
3. G. Pezzulo, L. W. Barsalou, A. Cangelosi, M. H. Fischer, K. McRae and M. J. Spivey, The mechanics of embodiment: a dialog on embodiment and computational modeling, *Frontiers in Psychology* **2** (2011).
4. J. Bonaiuto, E. Rosta and M. A. Arbib, Extending the mirror neuron system model, I, *Biological Cybernetics* **96**, 9 (2007).

5. 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 *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, eds. S. Ohlsson and R. Catrambone (Cognitive Science Society, Austin, TX, 2010).
6. T. Ziemke and S. Thill, Robots are not embodied! conceptions of embodiment and their implications for social human-robot interaction., in *Proceedings of Robo-Philosophy 2014: Sociable robots and the future of social relations*, (IOS Press BV, Amsterdam, NL, 2014).
7. J. R. Searle, Minds, brains, and programs, *Behavioral and Brain Sciences* **3**, 417 (9 1980).
8. S. Thill, H. Svensson and T. Ziemke, Modeling the development of goal-specificity in mirror neurons, *Cognitive Computation* **3**, 525 (2011).
9. B. Scassellati, How social robots will help us to diagnose, treat, and understand autism', *Robotics research* **552-563** (2007).
10. S. Thill, C. Pop, T. Belpaeme, T. Ziemke and B. Vanderborght, Robot-assisted therapy for autism spectrum disorders with (partially) autonomous control: Challenges and outlook, *Paladyn* **3**, 209 (2012).
11. S. Thill, P. E. Hemeren and M. Nilsson, The apparent intelligence of a system as a factor in situation awareness., in *Proceedings of the 4th IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support*, (San Antonio, TX, 2014).
12. B. Scassellati, Theory of mind for a humanoid robot, *Autonomous Robots* **12**, 13 (2002).
13. F. Varela, E. Thompson and E. Rosch, *The Embodied Mind* (MIT Press, Cambridge, MA, 1991).
14. J. Stewart, O. Gapenne and E. A. Di Paolo, *Enaction: Toward a New Paradigm for Cognitive Science* (MIT Press, 2010).
15. E. v. Glaserfeld, *Radical Constructivism* (RouteledgeFalmer, London, 1995).
16. R. Pfeifer, J. Bongard and S. Grand, *How the body shapes the way we think: a new view of intelligence* (MIT press, Cambridge, MA, 2007).
17. A. D. Wilson and S. Golonka, Embodied cognition is not what you think it is, *Frontiers in Psychology* **4** (2013).
18. B. Scassellati, H. Admoni and M. Matarić, Robots for use in autism research, *Annual Review of Biomedical Engineering* **14**, 275 (2012), PMID: 22577778.
19. S. J. Rogers, S. L. Hepburn, T. Stackhouse and E. Wehner, Imitation

performance in toddlers with autism and those with other developmental disorders, *Journal of Child Psychology and Psychiatry* **44**, 763 (2003).

20. C. Lord, M. Rutter, S. Goode, J. Heemsbergen, H. Jordan, L. Mawhood and E. Schopler, Autism diagnostic observation schedule: a standardized observation of communicative and social behavior, *Journal of Autism and Developmental Disorders* **19**, 185 (1989).
21. B. Scassellati, H. Admoni and M. Mataric, Robots for use in autism research, *Annual Review of Biomedical Engineering* **14**, 275 (2012).
22. D. J. Ricks and M. B. Colton, Trends and considerations in robot-assisted autism therapy, in *Proc. 1979 Int. Jt. Conf. Artificial IntelligenceIEEE International Conference on Robotics and Automation (ICRA)*, (Anchorage, AK, 2010).
23. B. Robins, F. Amirabdollahian, Z. Ji and K. Dautenhahn, Tactile interaction with a humanoid robot for children with autism: A case study analysis involving user requirements and results of an initial implementation, in *Proc. Paper presented at the IEEE International Workshop on Robot and Human Interactive Communication (ROMAN)*, (Viareggio, Italy, 2010).
24. H. Kozima, C. Nakagawa and Y. Yasuda, Interactive robots for communication-care: A case-study in autism therapy, in *Proc. IEEE International Workshop on Robot and Human Interactive Communication (ROMAN)*, (Nashville, TN, 2005).
25. B. Vanderborght, R. Simut, J. Saldien, C. Pop, A. S. Rusu, S. Pineta and D. O. David, Using the social robot probo as a social story telling agent for children with asd, *Interaction Studies* **13**, 348 (2012).
26. A. Tapus, A. Peca, A. Aly, C. P. andL. Jisa, S. Pintea and D. O. David, Children with autism social engagement in interaction with nao, an imitative robot — a series of single case experiments, *Interaction Studies* **13**, 315 (2012).
27. D. David, S. A. Matu and O. A. David, Robot-based psychotherapy: Concepts development, state of the art, and new directions, *International Journal of Cognitive Therapy* **7**, 192 (2014).
28. P. Oudeyer, F. Kaplan and V. Hafner, Intrinsic motivation systems for autonomous mental development, *IEEE Transactions on Evolutionary Computation* **11**, 265 (2007).
29. K. E. Merrick, A comparative study of value systems for self-motivated exploration and learning by robots, *IEEE Transactions on Autonomous Mental Development* **2**, 119 (June 2010).

30. D. Vernon, *Artific Cognitive Systems: A primer* (MIT Press, Cambridge, MA, 2014).
31. J. S. Kelso, *Dynamic patterns: The self-organization of brain and behavior* (MIT press, Cambridge, MA, 1997).
32. A. Clark, *Being there: Putting brain, body, and world together again* (MIT press, Cambridge, MA, 1997).
33. M. H. Bickhard, Autonomy, function, and representation, *Communication and Cognition-Artificial Intelligence* **17**, 111 (2000).
34. H. R. Maturana and F. J. Varela, *The tree of knowledge: The biological roots of human understanding.* (New Science Library/Shambhala Publications, 1987).
35. G. Schöner and J. A. S. Kelso, Dynamic pattern generation in behavioural and neural systems, *Science* **239**, 1513 (1988).
36. Y. Demiris and B. Khadhouri, Hierarchical attentive multiple models for execution and recognition (HAMMER), *Robotics and Autonomous Systems* **54**, 361 (2006).
37. Y. Demiris, L. Aziz-Zahdeh and J. Bonaiuto, Information processing in the mirror neuron system in primates and machines, *Neuroinformatics* **12**, 63 (2014).
38. D. Vernon, E. Billing, C. Costescu, D. David, P. Hemeren, S. Thill and T. Ziemke, Lightweight component-based software engineering in robotics: An architecture-first approach to system implementation and module integration, *Software: Practice and Experience* (submitted).
39. M. E. Harmon and S. S. Harmon, Reinforcement learning: a tutorial, *WL/AAFC, WPAFB Ohio* **45433** (1996).
40. X. Barandiaran, Behavioral adaptive autonomy. A milestone in the Alife route to AI?, in *Proceedings of the 9th International Conference on Artificial Life*, (MIT Press, Cambridge: MA, 2004).
41. T. Ziemke, On the role of emotion in biological and robotic autonomy, *BioSystems* **91**, 401 (2008).
42. T. Froese, N. Virgo and E. Izquierdo, Autonomy: a review and a reappraisal, in *Proceedings of the 9th European Conference on Artificial Life: Advances in Artificial Life*, ed. F. A. e Costa et al. (Springer, 2007).
43. A. Moreno, A. Etxeberria and J. Umerez, The autonomy of biological individuals and artificial models, *BioSystems* **91**, 309 (2008).
44. T. Froese and T. Ziemke, Enactive artificial intelligence: Investigating the systemic organization of life and mind, *Artificial Intelligence* **173**, 466 (2009).