The “Bottom-Up” Approach to Mental Life

A Commentary on Holk Cruse & Malte Schilling

Aaron Gutknecht

With their “bottom-up” approach, Holk Cruse and Malte Schilling present a highly intriguing perspective on those mental phenomena that have fascinated humankind since ancient times. Among them are those aspects of our inner lives that are at the same time most salient and yet most elusive: we are conscious beings with complex emotions, thinking and acting in pursuit of various goals. Starting with, from a biological point of view, very basic abilities, such as the ability to move and navigate in an unpredictable environment, Cruse & Schilling have developed, step-by-step, a robotic system with the ability to plan future actions and, to a limited extent, to verbally report on its own internal states. The authors then offer a compelling argument that their system exhibits aspects of various higher-level mental phenomena such as emotion, attention, intention, volition, and even consciousness.

The scientific investigation of the mind is faced with intricate problems at a very fundamental, methodological level. Not only is there a good deal of conceptual vagueness and uncertainty as to what the explananda precisely are, but it is also unclear what the best strategy might be for addressing the phenomena of interest. Cruse & Schilling's bio-robotic “bottom-up” approach is designed to provide answers to such questions. In this commentary, I begin, in the first section, by presenting the main ideas behind this approach as I understand them. In the second section, I turn to an examination of its scope and limits. Specifically, I will suggest a set of constraints on good explanations based on the bottom-up approach. What criteria do such explanations have to meet in order to be of real scientific value? I maintain that there are essentially three such criteria: biological plausibility, adequate matching criteria, and transparency. Finally, in the third section, I offer directions for future research, as Cruse & Schilling's bottom-up approach is well suited to provide new insights in the domain of social cognition and to explain its relation to phenomena such as language, emotion, and self.

Keywords
Bio-robotics | Bottom-up approach | Emergence | Evolution | Explanation | Mechanisms | Robotics | Social cognition

Commentator

Aaron Gutknecht
aaron-gutknecht@gmx.de
Johann Wolfgang Goethe-Universität
Frankfurt a. M., Germany

Target Authors

Holk Cruse
holk.cruse@uni-bielefeld.de
Universität Bielefeld
Bielefeld, Germany

Malte Schilling
malte.schilling@uni-bielefeld.de
Universität Bielefeld
Bielefeld, Germany

Editors

Thomas Metzinger
metzinger@uni-mainz.de
Johannes Gutenberg-Universität
Mainz, Germany

Jennifer M. Windt
jennifer.windt@monash.edu
Monash University
Melbourne, Australia

1 Biorobotics and the bottom-up approach to mental life

From my perspective, there are two basic ideas underlying the overall research strategy entertained by Cruse and Schilling. The first is that in order to understand a system and its properties, it has to be *reinvented* or *reconstructed* by the researcher. The second is that mental phenomena may arise as *emergent* properties via the interaction of low-level components of a system. I'd like to first provide an outline of these basic ideas and the underlying strategy as I understand them. In the next section, I will critically evaluate what types of questions the approach is best suited to answer, and what kind of problems it will likely face.

The first of the two main ideas is central to the research area of bio-robotics. If we are able to create an artificial system that exhibits the phenomena of interest, we should be a great deal closer to understanding how these phenomena come about in nature. In order for this approach to lead to valid conclusions, however, the process of reconstruction has to do justice to the systems we are seeking to understand. In the present context we are concerned, above all, with humans and other animals. This means that the way the artificial system achieves the desired results has to be biologically plausible, i.e., it has to be reasonable to assume that the capacities of the organism that we are trying to understand are really based on similar mechanisms. In this vein, Cruse & Schilling (this collection) are realising the basic reactive modules of their system in form of artificial neural networks that were inspired by biological research on, for instance, stick insects (Walknet) and desert ants (Navinet).

The second of the basic ideas derives its plausibility from an evolutionary perspective on psychological faculties. Emotion, attention, the ability to plan future actions, and any other "higher-level" capacities, including consciousness, did not arise suddenly from one generation to the next and independently of pre-existing, more fundamental abilities, such as the ability to control one's own body and respond adaptively to environmental stimuli. Rather these latter abilities and the interactions between the mechanisms responsible for them might well have been crucial for mental properties to evolve. From this perspective, the idea of reconstructing the evolutionary process by starting with basic reactive structures and examining how through the interaction of these structures unexpected properties might *emerge* seems very promising. Since humans also gradually evolved from simpler organisms, it is natural to assume that the same dependence between reactive structures and "higher-level" phenomena is present in our case as well. The investigation of this dependence might thus provide new insights into the mechanisms underlying human psychology.

But what does it mean exactly to say that a property *emerges* from basic structures? What is an emergent property? The philosophical controversies surrounding the concept of emergence date back over a hundred years and although usage of the term has become increasingly popular in recent years, among both philosophers and scientist, it can hardly be said to have one universal definition. Rather, there are numerous and varied interpretations, a fact which inevitably leads to confusion and misunderstanding (for a good overview see O'Connor & Wong 2012). It is thus vital to identify precisely what is meant by emergence in any particular case. Notwithstanding this inherent ambiguity, there seems, however, to be a shared idea behind much talk of emergent properties: this is the idea that as systems become increasingly complex they tend to exhibit certain higher-level properties, which are novel or unexpected given their simpler, lower-level, components.

Depending on how this claim is interpreted it can have more or less serious implications regarding the fundamental structure of nature, as well as the structure of science. In order to obtain a particularly strong and at the same time highly influential reading, it must be understood in a two-fold sense. First, as meaning that these properties cannot *even in principle* be predicted or explained on the basis of the lower-level properties of the system and, second, as indicating that such properties are associated with genuinely *new causal powers*, i.e., they make a real difference to the run of events and are not mere epiphenomena (for discussion see Kim 1999, 2006).[1] This kind of emergence could be called *strong emergence.*[2] Central to this conception is that emergent properties causally influence the simpler entities from whose organisation they emerge. This sort of causal influence is called "downward causation", as emergent properties are conceived as

[1] Such conceptions go back to thinkers such as Samuel Alexander, C. L. Morgan, and C. D. Broad, prominent figures in a philosophical movement, which came to be known as "British emergentism". The following discussion is, however, intended to illustrate the problematic nature of the concept of emergence and not to offer an analysis of the ideas of a particular philosophical school.

[2] It should be noted that the there is no universal definition of the term "strong emergence" in the current literature (for some alternative characterisations see Chalmers 2006; Bedau 1997; Yates 2013).

higher-level properties arising from certain lower-level properties and relations. Typically, it is assumed that what we find at the lowest level of this hierarchy are the properties and relations of fundamental physical particles. Given this assumption, the existence of emergent properties would entail that a complete description of the fundamental physical organisation of a system might still leave something out. The system might still have some properties that could not be predicted on the basis of such a description and could not be explained in terms of the organisation of its basic physical constituents. Moreover, because emergent properties are causally efficacious, knowledge of the basic physical components of a system and their behaviour may not be sufficient to predict the future evolution of the system. These considerations seem to lead to the conclusion that the meta-scientific thesis, according to which all phenomena can ultimately be explained by the fundamental laws of physics, would turn out to be false. If certain properties belonging to the domains of psychology, biology, or chemistry were emergent properties, these could not even in principle be captured by basic physics alone. All sciences dealing with genuinely emergent properties would remain completely autonomous, positing their own independent laws and explanations. Furthermore, since emergent properties have the ability to causally influence lower-level entities, the fundamental laws of physics would not even suffice to explain processes taking place at the *physical* level (see also Chalmers 2006).

These are substantial conclusions that could be met with some scepticism. They are also one of the reasons for the fierce controversy surrounding the concept of emergence. Furthermore, the condition that emergent properties are themselves causally efficacious and the general idea of “downward causation” leads to problems in and of itself. This is because there has to be a systematic relationship between emergent and lower-level properties, even though they are conceived as being distinct from another. Often this is expressed by saying that emergent properties are completely determined by lower-level properties and require them for their existence. In other words, if all lower-level properties of a system are fixed, its emergent properties are also fixed; and without any appropriate lower-level properties, a system cannot have emergent properties. If this weren’t the case, it would be unclear in what sense emergent properties *emerge from* lower-level ones (Kim 2006). If their relationship were completely coincidental, this would surely be an inappropriate description.

Based on this requirement, Kim (1999, 2006) has put forth an influential argument that the idea of “downward causation” is untenable. In summary, Kim’s basic argument is this: suppose an emergent property (let’s say a feeling of thirst) causes a lower-level property (e.g., a certain activation pattern N in the brain). If feeling thirsty is an emergent property, there have to be appropriate lower-level properties from which it emerges. Let’s call these the “emergence base” of feeling thirsty. Now, that feeling thirsty causes N means that there is a natural law that occurrences of feeling thirsty are always followed by occurrences of N (feeling thirsty is nomologically sufficient for N). But since occurrences of feeling thirsty are always accompanied by occurrences of its emergence base, it must also be true that occurrences of its emergence base are followed by occurrences of N. Therefore, if feeling thirsty causes N, its emergence base also causes N. But this makes feeling thirsty completely redundant as a cause of N. Its emergence base is completely sufficient to explain Ns occurrence, leaving the feeling of thirst as a mere epiphenomenon. Since this example can easily be generalised, one can conclude that there are no genuine cases of downward causation and hence no genuine emergent properties of the type presently under consideration.

In summary, it can be stated that emergence is a highly controversial concept—not only because of its inherent ambiguity, but also on account of certain varieties of emergentism that have substantial metaphysical and meta-scientific implications as well as a commitment to the problematic idea of downward causation. The crucial questions remaining now are whether Cruse & Schilling (this collection) provide a clear interpretation of the concept of emergence and whether it provokes the kind of controversy and criticism outlined above. What kind of emergence is involved in their claim that mental states might be construed as emergent properties? In fact, they provide two slightly different characterisations. According to the first, an emergent property is to be understood as a property of a whole system that cannot, *at first sight,* be traced back to the interactions of the systems components. Alternatively, one might say that we cannot, at first sight, predict the emergent properties of a complex system based on our knowledge of its parts and their interaction. Thus, we might be genuinely surprised that the system in question exhibits such properties. Emergence in this sense is sometimes called *weak emergence* (Chalmers 2006). If this is all that it means for a system to have emergent properties, few would raise serious objections. This sort of emergence is just a consequence of our limited knowledge and cognitive capacities and is relative to the judging subject: what might not be immediately predictable for one person might be just so for another. Emergentism, in this sense, has no far-reaching metaphysical or meta-scientific implications and is not committed to any sort of "downward causation".

Cruse & Schilling (this collection) provide a second, and equally unproblematic, definition of emergence that is specifically tailored for application in the context of robotics. According to that definition, a property of an artificially constructed system is emergent if it was not explicitly implemented by its designers. We might call this *implementational emergence.* This appears to be relatively independent of the sort of "weak" emergence I've just described. Even a property not explicitly implemented might be predictable without too much effort, whereas a property deliberately implemented might not be predictable, at least by persons lacking experience or competence. I think that most of the emergent properties Cruse & Schilling (this collection) attribute to their artificial system, reaCog, match both characterisations: they were neither explicitly implemented nor would we immediately expect or predict that reaCog would exhibit them. At the same time, the properties in question are highly interesting and are not simply insignificant side effects. This is important since, according to the definitions provided by Cruse & Schilling, the claim that an artificial system exhibits emergent properties is, *in and of itself,* not particularly notable. But this depends entirely on what the emergent properties in question precisely are. The finding that reaCog exhibits, in this way, aspects of psychological characteristics, such as emotion or attention and the ability to perform non-trivial body movements, are most certainly of considerable scientific significance. In conclusion, we may say that although the kind of emergentism advocated by Cruse & Schilling does not have the same far-reaching implications as the particularly demanding conception outlined above, it is nonetheless useful and philosophically interesting. This is because it functions as the basis of an intriguing approach to the study of psychological properties, which I shall now endeavour to describe.

Combining the idea of emergence with the idea, outlined above, that in order to understand a system and its properties, it has to be reinvented or reconstructed, we arrive at a fascinating research strategy. The first step consists in observing the behaviour of animals that, although lacking many of the sophisticated abilities with which humans are endowed, are nonetheless capable of flexibly controlling their bodies in order to cope with an unpredictable environment (such as stick insects, desert ants, and honey bees). Based on these observations one then develops a neural network model (e.g., Walknet or Navinet) designed to produce the behaviour observed in the first step. Next, this model is realised in an artificial system (either virtual or robotic) in order to examine to what extent the behaviour produced by the model matches the behaviour of the biological organism on which it is based. If it resembles it to a great extent, this can be taken as *prima facie* evidence that the mechanisms underlying the behaviour are the same for the animal and the robot. Different modules that are constructed in this way are then integrated into a holistic system. Further modules might be added step-by-step (e.g., Body Model, Attention-Controller, Word-Nets). The result is a complex system (in the present case “reaCog”) the behaviour and properties of which cannot be easily predicted even by its very own designers. The last, and most important step consists of examining whether the system shows characteristics that were not explicitly implemented but instead arise from the dynamic interactions of the system’s components. The most intriguing question in this context is, of course, whether the final system shows aspects of those phenomena that are constitutive of *having a mind*.

Although this is only a rough sketch of the methodology entertained by Cruse & Schilling (this collection), I hope I have captured the essential points sufficiently to proceed with an evaluation of its scope and the possible problems it might face. What kind of questions is the bottom-up approach best suited to answer? Which phenomena or processes can be addressed by research based on this approach? What considerations have to be taken into account in order for the presented research strategy to be successful? Are there any general constraints bio-robotic bottom-up explanations have to meet? As we shall see, the answers to these last two question are directly connected to two characteristics of the research strategy outlined in the previous paragraph: first, that it involves, at two points, a comparison of the behaviour of significantly different systems and, second, that it is specifically designed to discover emergent properties.

2 The bottom-up approach: Objectives, benefits and constraints

2.1 Mechanisms and the evolution of the mind

The most important aspect of the proposed approach is that it helps to elucidate the *mechanisms* underlying various mental properties. This is possible because many of the basic features of the control system reaCog are known. Using the words of Cruse & Schilling (this collection), it constitutes a “quantitatively defined system”. As all components are realised as artificial neural networks, all information about the number of neurons, the connection weights between them, and the way individual neurons process information is available. More importantly, however, the basic functional architecture of the system is well understood. Which modules are connected in which ways to other modules, how they receive their input, and what other parts of the system might be affected by their outputs does not have to be figured out by painstaking investigation—as is the case in biological research. Because these facts about reaCog are known, it is possible to provide detailed mechanism descriptions. In this way, reaCog’s ability to plan its future actions by internal simulation can be explained by reference to the interaction of its various sub-modules: a problem detector is activated when sensory input indicates that current behaviour will lead, if continued, to adverse effects for the system (e.g., falling over). This leads to the abortion of current behaviour and activation in the Spreading Activation Layer, which randomly excites the Winner-Takes-All network (WTA-net). After some time, the WTA-net adopts a relaxed state in which only one of its units is active. This active unit in turn stimulates its counterpart in the Motivation Unit Network, leading to activity of the corresponding reactive procedures. These provide motor output that can be redirected to the body model, which then simulates the execution of the proposed behaviour and predicts its likely consequences. If the system predicts that the problem will persist, the process of internal simulation goes on until a solution is found, which can then be used to control the actual movements of the system.

Explanations like these contain a lot of information about which functional subparts of a system are engaged during the exercise of the ability in question. In this particular case it makes clear how the ability to plan ahead, a cognitive ability, depends heavily on basic reactive structures that are designed to control specific leg movements as well as an internal model of the body. The same is true for various other capacities like attention and Theory of Mind. Thus, new insights into the mechanisms responsible for those phenomena in humans could be gained by considering how body models and motor control mechanisms are realised in our case and how these systems interact. In other words, the bottom-up approach may lead to new directions for future research concerning human psychology by suggesting how specific functional modules interact in order to bring about a particular target phenomenon. Whether this approach is tenable depends on the degree to which findings pertaining to the artificial system might legitimately be used to draw conclusions about human beings. I will propose a number of constraints to ensure that this condition is fulfilled below.

Another class of questions that a bottom-up strategy is well designed to answer has to do with the evolution of cognitive capacities: how did cognitive systems evolve from purely reactive systems? How did emotions, attention, or even consciousness arise? What are the natural precursors of these phenomena? Cruse & Schilling (this collection) show convincingly that no completely new neural modules are needed in order for such properties to occur. Rather, minor changes in the basic architecture might suffice to generate radical extensions of the abilities of a system. In this way, a reactive system with a body model can acquire the ability to plan ahead if it is able to disconnect its motor system from the physical body and instead send the motor signals to its internal body model. No novel “planning module” is needed. Already existing modules just have to become dissociable and can thus acquire new functions (Cruse 2003). In addition, the target paper suggests an answer to the question of the evolutionary function of cognition understood as the ability to plan ahead: it was the necessity of being able to control a complex body in a complex environment that made this ability highly valuable. Detecting problems by perception, finding innovative solutions by internal simulation and acting on them are capacities that are extremely advantageous for any organism possessing a body with a high number of redundant degrees of freedom (see Cruse 2003). This is in line with, and actually extends, the widespread assumption that the evolutionary function of cognition is to deal with environmental complexity (Godfrey-Smith 2002).

2.2 Constraints on bio-robotic bottom-up explanations

In the previous paragraph we saw that the framework Cruse & Schilling (this collection) present is well-equipped to give new insights into the underlying mechanisms of psychological phenomena and the evolution of cognition, as well as a promising approach to creating highly flexible and intelligent robots. There are, however, some problems the proposed strategy has to face, especially if the control structures become increasingly complex. I therefore want to suggest a set of three constraints on good bottom-up explanations of biological/psychological phenomena.

1. *Adequate matching criteria:*[3] At two points the research strategy described in section 1 involves a comparison between the behaviour of an artificial system on the one hand and a biological system on the other. First, this is the case in the development of neural network models of animal behaviour. In this context, the comparison is used to ascertain whether the proposed model of the mechanisms underlying certain capacities (e.g., walking) really reproduces the original behaviour of the animal (e.g., a stick insect). Second, there is a similar process of comparison involved in the application of psychological concepts to the complete system. At different points in their discussion, Cruse & Schilling (this collection) argue that their system has certain mental capacities because it exhibits behaviour (or would exhibit it if certain extensions were implemented) connected to those mental capacities in humans. So, for example, just as the performance of athletes might worsen if they consciously attend to what they are doing, the activation of the attention controller in reaCog can lead to poorer results compared to unimpeded execution of the reactive procedures.

   Both processes of comparison require criteria to identify when the behaviour of the artificial system and that of the biological system

[3] I credit this term to Datteri & Tamburrini 2007.

are relevantly similar, i.e., similar enough in order to provide evidence for the claim that similar mechanisms are at work in both cases or that the artificial system and the biological system share certain psychological characteristics (Datteri & Tamburrini 2007). The difficulty of finding such criteria increases the more the bodies of the compared systems differ. In some cases they might nonetheless be easy to find and relatively uncontroversial. This, however, is not always the case. For instance, in their discussion of emotions—and more specifically the emotion of happiness—, Cruse & Schilling (this collection) suggest that by increasing the threshold of the problem detector reaCog would take more risks, thus behaving similarly to humans when they are happy. Now, the question is whether the kind of risky behaviour exhibited by reaCog when the threshold of its problem detector is increased is the same kind of risky behaviour humans exhibit when they are happy. Only if this condition is fulfilled can the similarity be taken as evidence that reaCog shows aspects of the emotion of happiness.

2. *Biological plausibility*: Any proposed mechanism should be biologically plausible, i.e., it has to be reasonable to assume that the capacities of the organism that we are trying to understand are really based on such a mechanism. This can, at least to some degree, be ensured by trying to create similarities between the artificial and the biological organism on a basic structural level, for example by using artificial neural networks. Furthermore, it is necessary to decide how fine-grained the model should be. Should the model take brain structures, neurons, or subcellular elements as its basic building blocks? Should intracellular processes be neglected or are they important? The answer will of course always be relative to our particular epistemic goals. Finally, there are different options regarding the way artificial neurons process information, i.e., how they calculate their output value depending on the weighted sum of their inputs. All these factors might turn out to be important if the results are to be used to infer biological mechanisms. The requirement of biological plausibility shouldn't, however, be overemphasised. Cruse & Schilling (this collection) stress that they are not trying to present a realistic model of neuronal activity in living organisms. Hence, they are using biologically implausible, non-spiking artificial neurons as the basic elements of their architecture, while noting that some authors (referring to Singer 1995) have located the neural basis of consciousness in synchronously oscillating spikes. This, however, is not a weighty objection to the proposed approach since it is designed as a *functional approach*. The question is: how do different functional subsystems like a system for controlling the swing-movement of a leg, a system modelling the robot's body, and a system allowing for the selection of different internal states interact in order to produce certain emergent phenomena? Therefore, the concrete physical realisation of these subsystems is of only secondary importance.

3. *Transparency:*[4] Doubts about the strategy of using artificial systems in order to understand biological systems arise because even if we were to create an extremely intelligent robot, it would not necessarily help us to understand the mechanisms underlying its intelligence. Rather, we might be faced with yet another complex system whose workings we do not understand (Holland & Goodman 2003). Now, the approach Cruse & Schilling (this collection) present is specifically designed to discover emergent properties, i.e., properties that were not explicitly implemented. This means that there will be a high risk of finding properties in the complete system that cannot be readily provided with a clear-cut mechanistic explanation involving the cooperation of the system's components. Although the explanations of the occurrence of various psychological properties presented in the present paper are quite convincing, the

[4] The concept of transparency has a number of other well-established interpretations in the literature that should not be confused with the one at issue in the present context. These include, for example, "semantic" (Clark 1989) and "phenomenal" (Metzinger 2004) transparency.

bottom-up strategy might eventually exhaust its potential if the complexity of the system is further increased.

3 Future perspectives: The social insect

I would like to conclude by briefly proposing a perspective for future research based on the system reaCog. As presented, its ability to interact and cooperate with other agents is fairly restricted. At the same time, the pre-requisites of a broader social extension of the system seem to be in place. The present paper already shows how reaCog could be equipped with the capacities to recognize the behaviour of others and apply a Theory of Mind. In their 2011 paper, Cruse & Schilling further propose that by implementing a two-body model (a "We-model") reaCog might be capable of cooperative behaviour using shared goals. Integration and further expansion of such social capacities, and their application in an actual robot, seems promising considering the importance of social interaction in processes such as language acquisition and emotional regulation. Some have even suggested that the presence of other agents in the environment, or, in other words, dealing with social complexity, was a dominant factor in the evolution of sophisticated cognitive abilities (Humphrey 1976). Thus bio-robotic research in this direction might provide new insights into the mechanisms underlying such developmental and evolutionary processes. Moreover, a social extension of reaCog might eventually shed light on potential emergent phenomena *on a group level*, such as labour division, collective planning, social hierarchies and, most fundamentally, joint action coordination. What high-level social phenomena emerge when multiple bio-robotic systems like reaCog interact with each other?

Cruse and Schilling's system seems particularly well-suited to further illuminate motor theories of social cognition. According to such theories, the important social cognitive capacity of understanding another's actions is directly linked to mechanisms that are active when the observer performs similar actions Gallese et al. 2004; for criticism see Jacob & Jeannerod 2005). The underlying neural mechanism has come to be known as the mirror-neuron system. Furthermore, there is evidence that the mirror-neuron system plays a role in certain aspects of self-consciousness. For instance, Uddin (2007; see also Molnar-Szakacs & Uddin 2013) suggests that this is the case for representations of the physical self, and ascribes frontoparietal mirror-neuron areas an important function for self-recognition (especially the recognition of one's own face). As mirroring mechanisms can be integrated in reaCog as well, this opens the possibility of further investigating motor theories of social cognition and the relation between internal motor simulation and the self in a quantitatively defined system.

An ability that is highly important for human social interaction is the ability to communicate using language. At this point, the linguistic capacities of reaCog still seem quite inflexible and limited in scope. A highly interesting extension of this system would be to provide it with the means to learn words and their meanings by interaction with other agents. Some of the pre-requisites, like the ability to internally simulate the behaviour of others, could, as Cruse and Schilling argue, be implemented in reaCog by using its internal body-model to represent another agent. Robotic research in this direction was performed by Steels & Spranger (2009). Their artificial systems are capable of autonomously acquiring a simple language consisting of words for specific body postures. After learning is complete, the artificial agents are able to reliably assume body postures on verbal command by other agents. Since social learning has also been implicated in the process of concept formation (Steels 2002), the proposed extension might also foster our understanding of this intriguing phenomenon.

4 Conclusion

In conclusion it can be stated that Cruse & Schilling (this collection) present a highly fascinating research strategy that is well worth pursuing. The bottom-up approach can provide us with new insights regarding the functional mechanisms underlying psychological phenomena and their evolution. Although the notion of emergence is central to it, Cruse & Schilling (this collection) avoid the philosophical controversies surrounding this concept by interpreting it in a less demanding, yet interesting and useful way. There are, however, a number of constraints that explanations based on the bottom-up approach have to meet. First, since Cruse & Schilling's (this collection) strategy involves, at two points, a comparison between markedly different systems, criteria are needed according to which we can determine whether the two systems exhibit relevantly similar behaviour. Second, the structural architecture of the artificial system must have an adequate degree of biological plausibility. And finally, it has to be ensured that increasing the complexity of the system does not lead to the practical impossibility of elucidating the mechanisms underlying its emergent properties.

A promising next step for bottom-up research as presented by Cruse & Schilling (this collection) would be to take it to the level of social interaction. An extensive social extension of their system could shed light on a wide range of intriguing phenomena. Is it possible to discover emergent properties on a group level? In what precise way are mirroring mechanisms involved in social cognition? What role do such mechanisms play for the phenomenon of self-consciousness? What role do reactive structures and internal body-models play in the processes of language acquisition and comprehension? Of course this is only a small selection of the questions further bio-robotic research might contribute to answering. Cruse & Schilling (this collection) made clear that starting from the bottom is a strategy with enormous scientific significance. There is no doubt that this work will make an important contribution to a plethora of research projects in the future.

References

Bedau, M. A. (1997). Weak emergence. *Philosophical Perspectives*, *11* (s11), 375-399. 10.1111/0029-4624.31.s11.17

Chalmers, D. J. (2006). Strong and weak emergence. In P. Davies & P. Clayton (Ed.) *The re-emergence of emergence* (pp. 244-256). Oxford, UK: Oxford University Press.

Clark, A. (1989). *Microcognition: Philosophy, cognitive science, and parallel distributed processing.* Cambridge, MA: MIT Press.

Cruse, H. (2003). The evolution of cognition - a hypothesis. *Cognitive Science*, *27* (1), 135-155. 10.1207/s15516709cog2701_5

Cruse, H. & Schilling, M. (2011). From egocentric systems to systems allowing for theory of mind and mutualism. In T. Lenaerts, M. Giacobini, H. Bersini, P. Bourgine, M. Dorigo & R. Doursat (Ed.) *Advances in artificial life, ECAL 2011. Proceedings of the eleventh european conference on the synthesis and simulation of living systems* (pp. 185-192). Cambridge, MA: MIT Press.

Clark, A. & Schilling M. (2015). Mental states as emergent properties. In T. Metzinger & J. M. Windt (Ed.) *Open MIND* (pp. 1-39). Frankfurt a. M., GER: MIND Group.

Datteri, E. & Tamburrini, G. (2007). Biorobotic experiments for the discovery of biological mechanisms. *Philosophy of Science*, *74* (3), 409-430. 10.1073/pnas.1015390108

Gallese, V., Keysers, C. & Rizzolatti, G. (2004). A unifying view of the basis of social cognition. *Trends in Cognitive Sciences*, *8* (9), 396-403. 10.1016/j.tics.2004.07.002

Godfrey-Smith, P. (2002). Environmental complexity and the evolution of cognition. In R. Sternberg & J. Kaufman (Ed.) *The evolution of intelligence* (pp. 233-249). Hove, UK: Psychology Press.

Holland, O. & Goodman, R. (2003). Robots with internal models a route to machine consciousness? *Journal of Consciousness Studies*, *10* (4-5), 77-109.

Humphrey, N. K. (1976). The social function of intellect. In P. P. G. Bateson & R. A. Hinde (Ed.) *Growing point in ethology* (pp. 303-317). Cambridge, UK: Cambridge University Press.

Jacob, P. & Jeannerod, M. (2005). The motor theory of social cognition: A critique. *Trends in Cognitive Sciences*, *9* (1), 21-25.

Kim, J. (1999). Making sense of emergence. *Philosophical studies*, *95* (1), 3-36. 10.1023/A:1004563122154

——— (2006). Emergence: Core ideas and issues. *Synthese*, *151* (3), 547-559. 10.1093/acprof:oso/9780199585878.001.0001

Metzinger, T. (2003). *Being no one: The self-model theory of subjectivity.* MIT Press.

Molnar-Szakacs, I. & Uddin, L. Q. (2013). The emergent self: How distributed neural networks support self-representation. *Handbook of neurosociology* (pp. 167-182). Dordrecht, NL: Springer.

O'Connor, T. & Wong, H. Y. (2012). Emergent properties. *Stanford Encyclopedia of Philosophy.* http://plato.stanford.edu/entries/properties-emergent/

Singer, W. & Gray, C. M. (1995). Visual feature integration and the temporal correlation hypothesis. *Annual review of neuroscience*, *18* (1), 555-586. 10.1146/annurev.ne.18.030195.003011

Steels, L. & Kaplan, F. (2002). AIBO's first words: The social learning of language and meaning. *Evolution of Communication*, *4* (1), 3-32. 10.1075/eoc.4.1.03ste

Steels, L. & Spranger, M. (2009). How experience of the body shapes language about space. In M. Kaufmann (Ed.) *IJCAI'09: Proceedings of the 21st international joint conference on Artifical intelligence*. San Francisco, CA: Morgan Kaufmann.

Uddin, L. Q., Iacoboni, M., Lange, C. & Keenan, J. P. (2007). The self and social cognition: The role of cortical midline structures and mirror neurons. *Trends in Cognitive Sciences*, *11* (4), 153-157. 10.1016/j.tics.2007.01.001

Yates, D. (2013). Emergence. In H. Pashler (Ed.) *Encyclopaedia of the Mind* (pp. 283-287). San Diego, CA: SAGE Reference.