Uwe Saint-Mont (Nordhausen Univ. of Applied Sciences)

Roads to Consciousness:

Crucial steps in mental development

Abstract

For a long time, philosophers have considered the conundrums of consciousness, self-awareness and free will. Much more recently, scientists have joined in and begun to unravel the secrets of mind. Biologists, physicians and psychologists, studying the human brain, but also physicists, engineers, and computer scientists, working on organizational principles of intelligent information processing systems, have contributed to the subject.

This contribution explains several "roads to self-awareness", all of them based on the natural sciences. The first one follows our bio-psychological evolution. The second road starts with the engineer's point of view and mainly builds on information science and technology, in particular robotics. The third road taken is the most abstract; it exploits complex dynamic systems and their emergent properties.

Despite their different origins and methods, these lines of investigation converge. That is, the findings of various fields can be combined into a unified theory of mind and self-awareness, which is the main purpose of this paper.

This overall synthesis suggests that the mind results from a multi-hierarchical organizational structure, and self-reflexive flows of information in embodied systems. In addition to this, stable self-awareness appears spontaneously in sufficiently complex robots, when the system's capability of describing itself crosses the level of conceptually clear information processing (thinking).

As an application, one obtains a number of construction principles for mentally developing systems that are explained towards the end of this contribution.

Keywords

self-consciousness, self-awareness, free will, dynamic systems, hierarchical systems, language

*PsycINFO Classification:*2100; 2600

1) A psychological theory of self-awareness

"What," said I, "does the puny creature mean by 'it'?" "He means himself," said the Sphere: "have you not noticed before now, that babies and babyish people who cannot distinguish themselves from the world, speak of themselves in the Third Person?" (Abbott 1884)

Following the path of biological evolution, Saint-Mont (2001) delineated a concise theory of personal self-awareness. It is the aim of this section to expand and improve upon these ideas.

Step 1: Information processing. The most important property of (animal) nervous systems is that they process information. The incoming information, the input of the system, stems from the outside world. This input may be stored and internally processed in many ways. But if it is to be of any value for the animal or human being, these stimuli finally have to be transformed into reasonable motor actions, that is, behavioural responses that are adequate with respect to the momentary situation. In a nutshell, stimuli are processed somehow in order to reach some behavioural response:

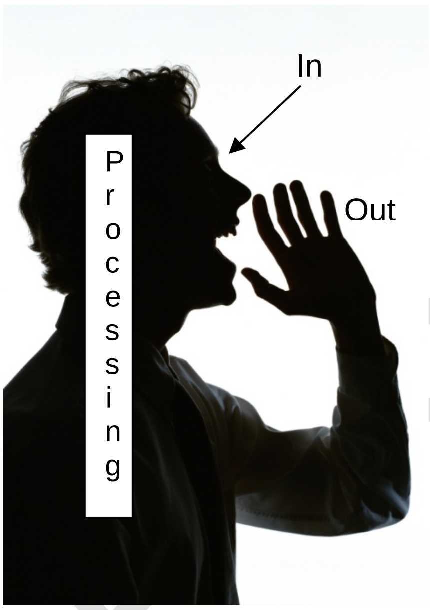

Illustration 1 (Linear Information processing: Stimulus-Organism-Response)

Step 2: Representation. In order to do so, the brain needs to represent relevant parts of the external world. These representations are based on the input provided by the sense organs, and every sense organ is associated with typical units or items: touches, smells, tastes, sounds, and of course, images. The items may be stored and retrieved from an internal memory or be evoked by some external stimulus. But these are details. The point is, that, altogether, they constitute a (possibly very crude) model of the world, which, despite all the interrelations amongst the items, must be based heavily on the sensory input if it is to be of any value to the individual. For human beings, images are by far the most important representations.

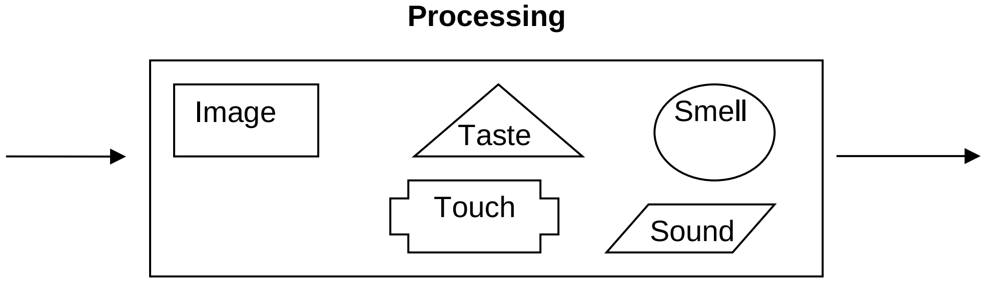

Illustration 2 (Various internal representations, e.g., of a single external object)

Step 3: Representation of own body. There are not just sensory impressions of the external objects. The mental realm also contains representation(s) of the animal's physical body or body parts. These special items are rather easy to obtain, since all input information is associated with sense organs that are embedded in the physical body.

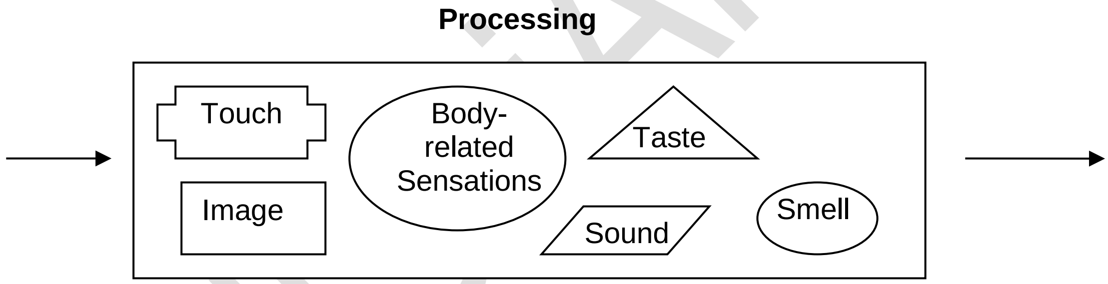

Illustration 3 (Internal representation of body parts or body-related phenomena)

Step 4: Circularity. On the one hand, it is the body which is receiving external information. On the other hand, the body is acting in the outside world. Therefore, things start to become circular here: The individual takes action in the real world, causing some change there, which subsequently - within hours, minutes, or seconds - may have some noticeable consequence. In other words, for every animal but also every human being, anything takes place around a centre which is the personal body (being agent as well as observer).

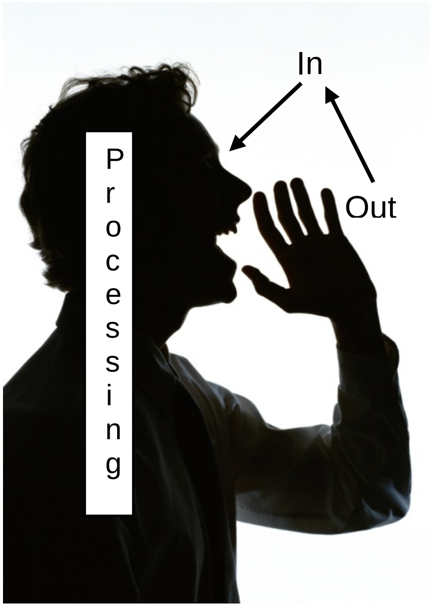

Illustration 4 (Circular Information processing: Stimulus-Organism-Response – Stimulus – Organism – Response, etc.)

Step 5: Self-perception. If action and perception are closely related, e.g., if the animal's (own) body is acting and - almost at the same time - the animal perceives that the body (located in the centre of activity, and being of paramount importance) is in motion, it is a small step to assume that the animal "notices" itself. That is, it observes that there is something special about this body; that a distinction should be made between oneself and the rest of the world. As human beings rely mostly on the visual sense, the perceived image of the own body is by far the most important representation of oneself.

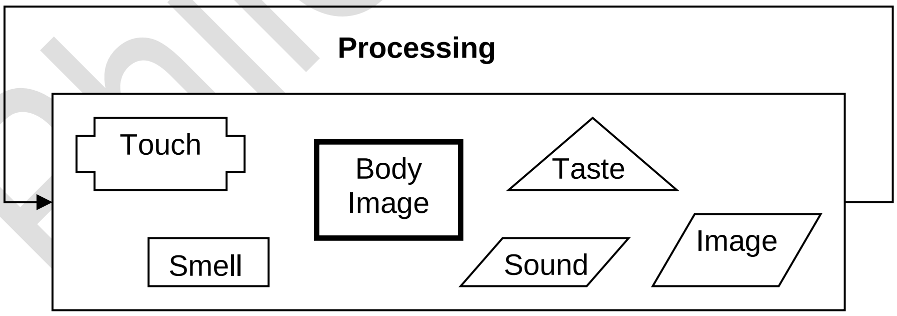

Illustration 5 (An accentuated image/representation of the body emerges)

Step 6: Language. The crucial innovation of homo sapiens is an effective, omnipresent language. With the naming of objects, the verbal description of facts and the narrative planning of actions, a second, language-based internal representation (i.e., model) of the real world evolves. Although the verbal model is strongly connected with the first (mainly visual) representation of the world, especially via concepts, the individual has *two* distinct ways to realize things. Typically, two representations of one and the same fact are available – an *image* and a *name*. A concept is just this: An image (or sensory impression, more generally) plus a corresponding verbal description.

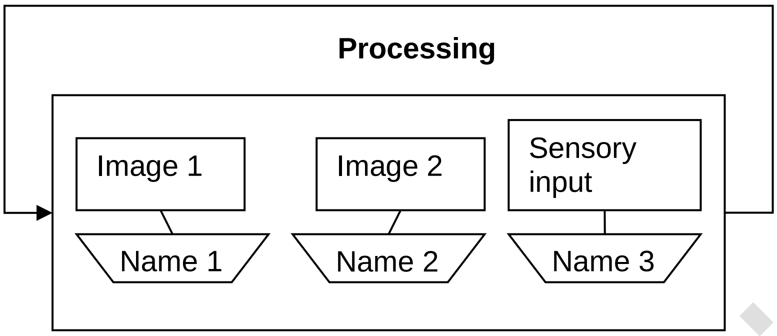

Illustration 6 (Concept formation. Concept = sensory input labelled with a name)

Step 7: A special image and a peculiar name. In particular, perception and language yield two distinct representations of the subject. There is a nonverbal and a verbal description of oneself available: the image of the body – which already has had an accentuated position - and its (specific) name. Body and name are not like all the other objects, there is something special about them, for word and image – both - represent the individual.

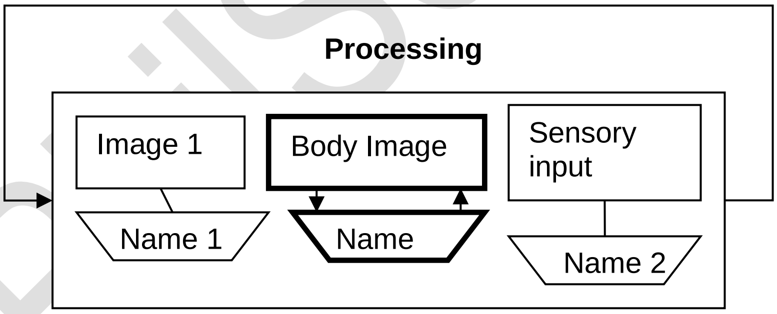

Illustration 7 (The Self-Concept = Body Image & corresponding name)

Step 8: Self-Awareness. Only humans have two distinct ways of description and thus an "internal mirror": When processing words and sensations, the body image is reflected in its corresponding name and the name has a counterpart in the corresponding body image. Something extraordinary happens when these parts melt into a unit. Image and name combined constitute a concept of oneself (or one's self, respectively). The entity that emerges is a self-concept of the individual, and the individual becomes self-conscious, i.e., fully aware of its position in the world. Identifying the body's name with the body's image is the crucial step that yields the self: an entity in the middle of everything, right in the centre of action and sensation, yet distinctively different from anything else in the world, and of paramount importance.

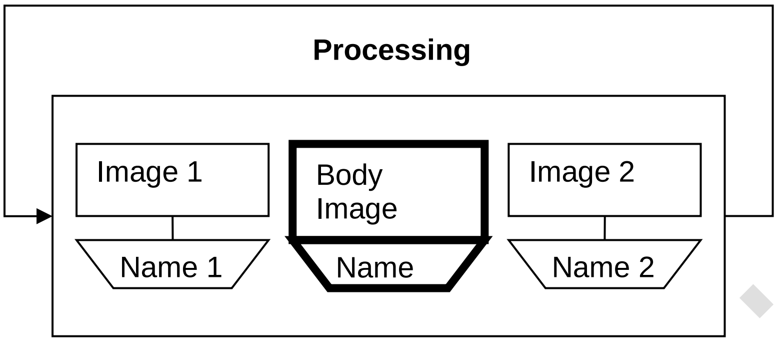

Illustration 8 (In the mental realm, clear self-awareness is the result of a crisp, stable distinction between the body concept and anything else)

In other words: Self-awareness emerges when we learn to draw a clear-cut cognitive line between us and the rest of the (perceived) world. It is this *permanent, stable distinction* which, according to Saint-Mont (2001), constitutes personal awareness. We are self-conscious beings, aware of our individuality, because of a stable contrast between a world "out there" and ourselves (or "our selves", respectively)

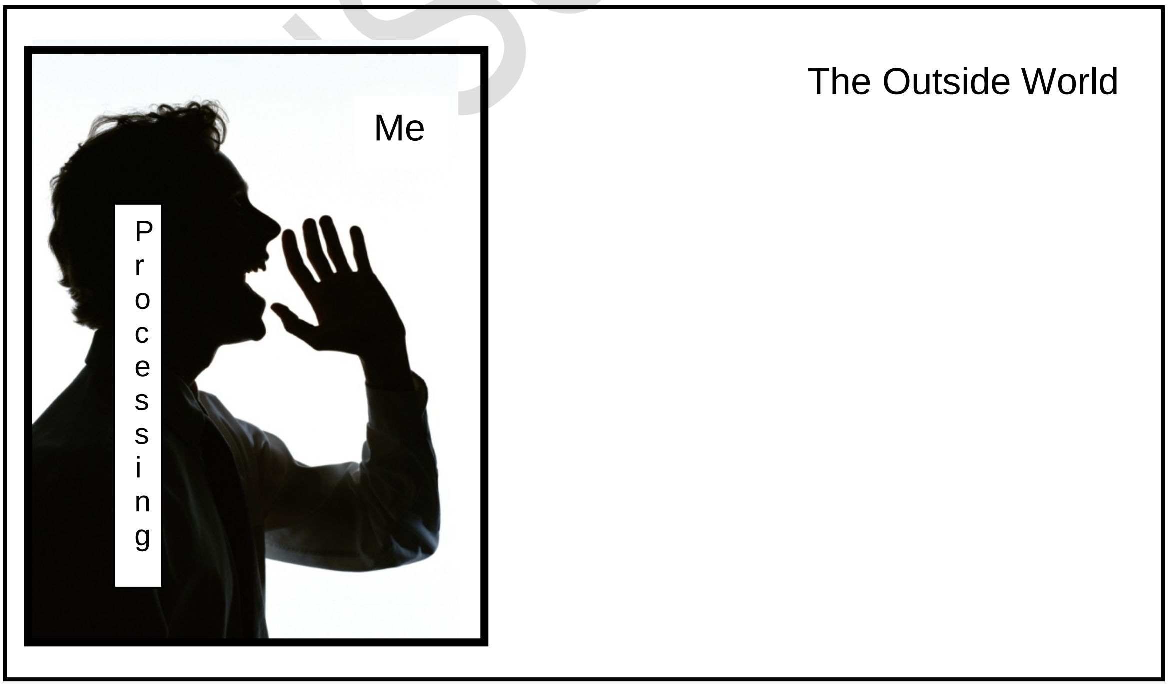

Illustration 9 (Given a person and their environment, conceptually clear self-awareness distinguishes oneself - one's self - from anything else)

2) Psycho-biological reasoning and evidence

It's the brain in a body in a world that matters (Smith 2009)

In a sense, much of the reasoning of the first section is well-known. However, it is one thing to be aware of some relevant components and steps to be taken. It is quite another to assemble these pieces into a logically-sound blueprint of an extremely complex machine. To this end, we next embed Saint-Mont's train of thought into a more general conceptual framework.

Step 1: Information processing. The S-O-R-paradigm is a classic. Introduced by behaviouristic psychology one hundred years ago (Watson 1913), it has since been adopted by many fields. Since a classic computational device reads input, processes it, and produces an output, "S-O-R" ("Input-Process-Output") is also the most fundamental model of information processing in the computer sciences.

Step 2: Representation. It is rather obvious that the sensory stimuli reaching our sense organs are processed and transmitted to the brain. Thus, strictly speaking, we do not know the external world as it really is. All we are aware of are certain "internal" impressions that (seem to) stem from our eyes (vision), ear(s), hands (haptic), etc. Of course, ever since Plato's parable of the cave, philosophers have been discussing this situation at length. Contemporary scientists face a similar problem: If a robot is to accomplish some task, it first of all needs to be informed about (relevant aspects) of its environment. It is difficult to hit a target if one is blind and deaf. Thus a crucial question becomes how to represent or "model" the external world in the robot.

Step 2a: Integration. Now, a certain crucial step seems to be missing in Saint-Mont's account. Given some external object, this object is perceived by various sense organs. An observer sees a cup of coffee, smells its odour, recognises its temperature, and tastes the characteristic flavour. All these bits of information, transmitted by various channels – to use modern jargon – need to be integrated into a comprehensive and single impression: a fine cup of coffee, or, more precisely, the cup of coffee as you perceive it. Gestalt psychology stressed the necessity of integration with respect to visual perception; but also many contemporary authors emphasize that diverse sensory impressions need to be combined. Damasio (2010) calls the integrated chunks of information "maps" (thus an "image" in Saint-Mont's words is a "visual map" in Damasio's terminology), and in a chapter entitled "Putting it together", he highlights the role of the brain stem in this endeavour.

Step 3: Representation of own body. Given this, there is (at least, at first) not much special about one's own body. Various sensory impressions are combined into one mental entity, typically called a body image (cf. Gallagher 2006, De Preester and Knockaert 2005). In other words, any animal or robot, possessing a body and being equipped with sensory organs is able to perceive its own body and may thus form a comprehensive body image. But although, in a sense, it is a map like any other map, there is something peculiar about it. You perceive your own body from a particular unique perspective. Owing to your position and its lookout, you see at least parts of it, e.g., your arms, chest, belly, legs and feet. Recognizing your own voice may even be easier than listening to others. Moreover, as already mentioned, this – your - (integrated) body map takes centre stage.

Step 4: Circularity. Moreover, the flow of information changes drastically, which makes completely new "emergent" phenomena likely. The crux of the Chinese room argument (Searle 1980) is that the room (or anybody in it) does not know what is really going on: Received input is just transformed into output in a perfect way. The machine has no concept, no map or token for itself. With the information processing going on within the body, and the body being a major player, the situation is completely different. The information flow is no longer linear, but circular; i.e., output may become input:

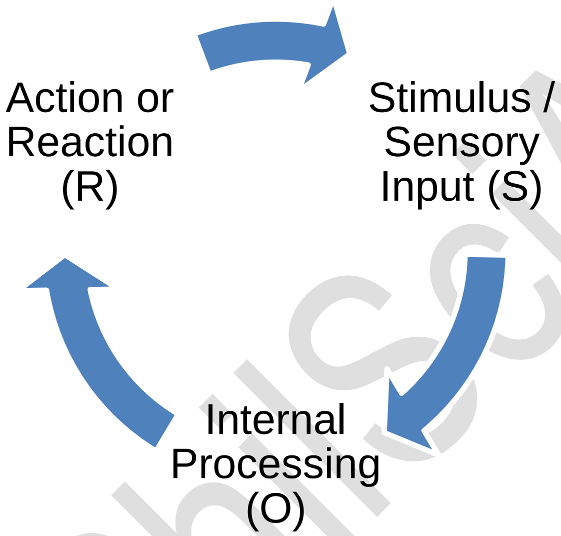

Illustration 10 (Circular Information Processing – the sensorimotor loop)

For this situation, computer scientists coined the term "embodied cognition", and placed it in stark contrast to traditional artificial intelligence: "Instead of emphasizing formal operations on abstract symbols, the new approach…foregrounds the fact that cognition is a highly *embodied* or *situated activity*…, and suggests that thinking beings ought therefore be considered first and foremost as acting beings" (Anderson 2003, p. 91, italics in the original).

"Grounded cognition" is also a prominent new concept in psychology: "[It] rejects traditional views that cognition is computation on amodal symbols in a modular system, independent of the brain's modal system for perception, action, and introspection" (Barsalou 2008).

With the “senorimotor loop” (Der & Martius 2012) or “perception-action” loop (Shapiro 2010) in place, perceptions are always related to the body, which subsequently may take suitable, i.e., input-dependent actions (S-O-R). Starting with motor actions, they and their consequences can be perceived and may have some impact on the body (R-S-O). Finally, it is only the body that can take action and perceive what has happened (O-R-S).

Various iterations of the loop (S-O-R-S-O-R-...) reveal that there is something special about the body (O): It does not just take centre stage with respect to perceiving (it has a unique perspective), it also takes centre stage with respect to acting – its “effectors”, i.e., its hands, feet etc. change the environment of the body. Using this loop effectively means learning what consequences an action has for the body, i.e., an inept beginner may evolve into an adept master. Subsequently the processing within the body may choose a certain action in order to provoke a certain effect on the body.

Accordingly, science nowadays distinguishes between the (integrated) body image which is mainly sensory, the (complete) body schema which is both sensory and motoric, and agency which is mainly motoric (De Preester and Knockaert 2005). But, of course, all these entities are intensely linked. In humans, precise eye-hand coordination is nothing but a tight feedback loop of the above kind: We embark on a certain action, observe intermediate results, may thus alter our (re)actions, until, finally, we obtain a desired result involving some external object or one’s own body. Obviously, there are many feedback loops around, involving other effectors and sense modalities. Moreover, internal feedback loops within the brain seem very likely. We may thus simulate a certain action and anticipate its results without actually performing it in the real world.

Step 5: Self-perception. Due to the circular situation, the body plays a double role. It is a (rather) passive object being perceived and (its map) being processed. At the same time, it is the active subject, doing this processing and acting in the external world. On the one hand, just like any other object, the body is perceived by the sensory organs, and represented by a cognitive, i.e., internal map. On the other hand, the body’s nervous system, in particular its brain, is doing all this processing. That’s a very peculiar and unique property, distinguishing the body and its mental activity from all the other objects around.

All perceptual as well as motoric information is linked to the brain. More precisely: Within the flow of information coming from the senses and finally resulting in motor actions resides “central processing”, mainly going on in the brain. The brain’s thoughts are an integral part of the complete situation, or – rather - a pivotal element of/in the sensorimotor loop. Now, if feedback is strong and rather instantaneous, i.e., if the various perception-action loops are tight and numerous, the body and its mental processes can hardly escape their own presence. Thus a straightforward question arises: How much do they “notice” their own activities? Or, to put the question slightly differently: How much does a body endowed with an information processing unit understand about its position, and the role it plays in the above situation? In particular, is it able to distinguish between private and external, self vs. context?

Obviously, the answer to this question forms a **continuum**, the continuum of self-awareness. One extreme consists of beings (be they living or artificial) without the slightest idea about themselves. The other extreme shows up in healthy, grown-up humans who know exactly where they are and what they are doing. In between seem to be animals (and perhaps robots) that - more or less - understand their situatedness. Depending on their “equipment” (both mentally and physically) they come more or less closer to the “human end” of the continuum of self-awareness. However, since complex life forms originated more than half a billion years ago and humans are the only conscious species we know, it also seems to be very difficult to overcome obscurity, and to reach the “enlightened” endpoint. Thus a straightforward question is which powerful tool(s) enabled man to get there.

Step 6: Language. To cut a long story short, many scientists and philosophers think that language makes the difference (Hauser et al. 2002). That is, the single most important difference between man and animals (even the most developed ones) consists in our exclusive language skills. On the one hand, those animals considered closest to us are those with remarkable language skills, in particular certain primates, whales and birds. On the other hand, it has been reported that people who learned language late in their lives refer to themselves as some “phantom” that existed before. See, for example, the “extraordinary mind of Helen Keller” in Donald (2002, Chapter 6), and Schaller (1991).

With language comes conceptual precision and clarity. One may systematically name objects and delineate a situation. Whereas maps are located on the perceptual “side” of the sensorimotor loop, speech production is an active feature, rather located on the motor “side”. Thus language immensely helps in describing, analysing, and moving in the world we inhabit. It makes way for a deeper understanding of our natural and social environment, our place in it, and our personal characteristics - be they external (such as the expression on my face) or internal (like the mood I am in).

Step 7: A special image and a peculiar name. Since the purpose of our most important sense organs is to collect information about the outside world, these organs are directed away from us. However, when we look into a mirror the view is thrown back. I see my own body and my face, a peculiar part of my body, distinguishing myself from everybody else in the world. It is surely no coincidence that, with the help of this strong and immediate visual feedback, at least some animals of a few species are able to recognize themselves (Gallup 1970). They notice that the body and face they encounter are something special, that this impression differs from all other objects. (For an up-to-date overview of the species passing the “mirror test” cf. the same-named entry in the English wikipedia.)

In other words, with the help of the mirror’s immediate feedback, there is an additional loop, and the mental processes in an animal’s body are able to distinguish between own and alien. That is, in front of a mirror, the animal - or rather its information processing - is able to draw a (cognitive) line between its individual existence and the rest of the world. In this sense, the mirror acts as a catalyst towards self-awareness. However, if these animals look in a different direction, the loop is gone and they seem to lose their fundamental insight almost immediately.

With our sense organs intact, humans - but also our cousins in the animal kingdom - perceive a rich model of the real world. That is, without any effort, we all observe what is going on around us. Homo sapiens, however, is the only species that is able to describe the situation in a second, completely different way. The crucial point in Saint-Mont’s account seems to be that humans with a versatile, powerful language system possess a second (verbal) tier and thus a fully functional “internal mirror”.

Already within the perceptual model alone, there is a special entity: the body. Since it is the central unit, all action taking place around it, and observation being directed towards it, it plays the chief part. In the world of language all kind of objects and phenomena are given names. With the help of these concepts, one can speak about everything in a rather concise way. Here, too, evolves a concept with a special meaning. It is the concept that names the individual’s body and anything directly related to it, such as the visual appearance, the sound of the voice, actions initiated by the body, and internal states.

So, there is something special about the concept “I”. This concept is very different from all other concepts. It describes the centre of existence, the source of actions and the spot where all perceptual input converges. In the verbal realm, everything that is going on “revolves” around this crucial token. “I” has the same status that the body map possesses in the non-verbal world model.

Step 8: Self-Awareness. From the very beginning of verbal utterances, the word “I” and the body map are close. For what is the meaning of the latter concept? Its semantics are always very much related to the body’s perspective, its parts, its actions, and, last but not least, its information processing going on within the very head of the body.

Animals of all species possess at most one elaborated perceptual model. Some, like dolphins, use complex communication systems in addition. However, homo sapiens seem to be the only species with two very rich systems of description. Although they are closely related, they exist in their own right, and there is a pronounced difference between the world as we perceive it and the world we are talking about. An image and a concept are completely different chunks of information.

Animals need the help of an external mirror to encounter themselves. With a rich perceptual as well as a sophisticated verbal model, that's different. Each of these models can serve as an "internal mirror" to the other. Moreover, both include a marked representation of oneself: the body map on the one hand, and a peculiar word for the individual on the other. In other words, the reflection of the concept "I" is the body map; and the body image is represented by the term "I" in the realm of language.

This internal feedback loop is immediate, tight and strong. In order to reach a clear understanding of oneself, all that is still needed, is the identification of word and map, of a peculiar name and its visual image. When these two objects fuse, a comprehensive concept emerges, encompassing all properties belonging to the extraordinary entity right in the middle of everything that is going on. On the one hand the chunk of information standing for oneself is body-related (all we sense, feel, think and do at a certain moment in time); on the other hand it is a clear-cut, precise concept. On the one hand it is passive/receptive (e.g., the image we see in a mirror upon opening our eyes), on the other hand it is active/motorial (e.g., planning, volition, and taking action). Thus a personal self is born, one's very identity established.

One could also say, and that is Saint-Mont's emphasis, that the personal self comes into existence due to a permanent, stable distinction between oneself (or: one's self) and anything else. Learning the distinction between own and alien is also considered crucial in developmental psychology. Rochat (2003) writes [italics in the original]: "Until the middle of the second year when linguistic and symbolic competencies start to play a major role in the psychic life of children, self-awareness remains *implicit*. It is expressed in perception and action, not yet expressed via symbolic means such as words. Prior to approximately 14–18 months there is yet no clear evidence that the children perceive traces of themselves, as *standing for* themselves, only themselves, and no one else, such as the little footprints they might leave in the mud or the image they see in the mirror." He calls the crucial step "identification": "At this level, the individual manifests recognition, the fact that what is in the mirror is 'Me,' not another individual staring and shadowing the self."

Thus, yet another way to express this step is to say that we are no longer strangers to ourselves. Instead of looking at ourselves from the outside, always with a certain distance, we understand that we are the one perceiving and the one being perceived, that we are acting and, at the same time, observing what we are doing. Listening while we are speaking triggers the insights that (1) speaker and listener are the same person, and (2) that we are trying to understand our own words, not somebody else's. In a nutshell, we come into our own, distinguish our identity from all others, and recognize our self (or: ourself) in a clear, conceptual way.

This step is crucial, since we thereby reach the right end of the mental continuum. The centre of all perception and activity leaves vagueness behind, the human agent reaches a completely new level of insight – the cognitive level. ("Cognitive" meant in a narrow verbal-conceptual sense; as opposed to imprecise emotions, multifaceted perceptions, not integrated lines of information processing, and a fuzzy perforated boundary between "inside" and "outside", oneself and others.) That is, man finally understands the basic setup of the game and his distinguished role in it. Upon integrating all individual-related information into one conceptual entity, he forms an identity. Equivalently, one could say that upon establishing a stable frontier between own and alien the personal identity is assembled. The frontier is mainly conceptual and it is fundamental to all cognitive processes, since it creates a precisely defined "me". So, equivocally speaking, "self comes to mind" (Damasio 2012), whereas, before, there was just "the feeling of what happens" (Damasio 1999).

Step 9: Major consequences

Steps 1-8 seem to be straightforward. In particular, I described the last, crucial step as if it appeared all of a sudden, in a certain moment of "enlightenment". Of course, in a certain sense, establishing a self is like picking a ripe fruit from the tree of knowledge. However, in the physical world, developments take time. First, it is well-known that newborns need many months to develop the mental capacities necessary, in particular language skills, for them to finally reach their selves. Second, the crucial insight may appear all of sudden, but it may also be forgotten in a minute. Therefore, third, it takes months until the "Me-But-Not-Me dilemma" (Rochat 2003) is finally dissolved, i.e., a stable individual identity endowed with clear self-awareness is established.

Fascinating as these developmental details may be, much more important is the reorganization of the psychological arena that occurs subsequently. With a perceiving and acting agent in the middle, perfectly aware of its position in the world, the information flow is altered dramatically. The self "coming to mind" triggers a *fundamental reorganization of the mental landscape* and completely new effects emerge:

1. Thinking becomes conceptually clear
2. Planning is thus rendered deeper and more complex
3. Subsequent actions taken are well-aimed
4. Attention steers the sensors towards the most interesting phenomena
5. Memory can become more selective, saving important information
6. An autobiographical memory occurs
7. The individual may explore its feelings, traits and other inner features
8. A sense of property emerges (all things that belong to me, but not to others)

Altogether, step by step, these abilities potentiate the individual's reach, and, make no mistake, it is the developing agent that is actively extending its force. There is an owner who learns to handle the mental and physical tools available, and, in the end, can apply them as he pleases. Fortunately for us, it turns out that human brains are embedded into a versatile body with an appropriate size, and well-functioning in almost any natural environment, on land as well as in water. Even more important is the fact that we are able to design, to change our environment with the help of very sophisticated effectors: our hands (Wilson 1998). That's a lucky coincidence, since self-awareness could also be "locked" in a body tailored to a narrow ecological niche; just suppose you were a raven, a dolphin or an elephant...

On their own, individuals can survive: they can assemble tools and equipment, hunt, produce clothing, and may even build a hut. However, many hands, working together, are needed to piece together megalithic sites, pyramids, or walls stretching thousands of miles. With the help of language, writing and many more cultural techniques, man could organize larger, work-sharing, stable groups which turned out to be the nucleus of complex societies. Nowadays, this historic quest seems to be cumulating into one truly global culture.

Impressive as all these steps are, we have omitted the single most important one, occurring quite early and adding to our cognitive life a completely new dimension. This single most important personal insight is the detection of time. Unlike all animal species, we are not just living in three-dimensional space: Looking back, we see that we were younger, with people telling stories about our birth, when our subjective life started. Looking ahead, however, each and every one of us has to concede that he is growing older, until finally, his life is over. Understanding the past, and foreseeing at least a part of the future is an invaluable gift, it deepens and widens our consciousness immensely. However, this gift inevitably comes with knowledge about our inevitable fate. Each and every one of us must foresee and thus face the fact of death, i.e., a limited existence in time.

Consciousness has been a great invention, perhaps it has been the most powerful innovation ever since the Cambrian revolution, reaching a completely new level of insight and complexity, shaping much of us (our culture, and history), and altering the face of the planet. Nevertheless, self-awareness - in essence being a mental borderline - comes with restrictions and limits: Opening one's eyes in the middle of the night won't make the sun shine, since the sensory system and the view of the world it provides are not affected by cognitions.

In more general terms, consciousness is a higher-level mental process with a certain influence. However, this process neither understands nor controls the psycho-physiological machinery completely. Freud and many others have pointed out that major mental tokens, like motivation, emotions, drive, pleasure or pain are beyond its reach. There is voluntary as well as involuntary motor function. Moreover, everybody is born with a certain set of physical and mental properties. These properties constitute basic conditions under which our lives evolve. (It really makes a difference if one is blind or keen-eyed, emotionally stable or fragile, can move his limbs or not.) Although we are able to talk about almost anything, we are clearly aware of much and we are able to change some conditions, there are always many boundary conditions that we may neither overview, nor understand, nor able to alter them.

So, finally, there we are: A well-defined identity with a distinct personality, precisely knowing where it is located in time and "space" (the latter being physical and social). In all the fields described our boundaries have widened. However, inevitably, our psychological life is tied to a particular body. Instead of being like a "spirit hovering above the waters", we are "embedded intelligence", inseparably linked to some physical entity, to the extent of being this body's agent.

3) The engineer's perspective

What I cannot create, I do not understand (Feynman 1988)

Although the arguments so far have focused on human onto- and phylogenesis, they can also be understood in much more general terms. In a sense, it would be very surprising if homo sapiens could be the only self-conscious system or if biological details were decisive. Thus, although the above steps mainly used psychological and biological concepts, given a detailed enough blueprint, containing all crucial technical details, one should be able to construct conscious systems. Planes fly since we have understood the physics, not because we imitate birds perfectly.

Programmable robots

Let us therefore look at the problem of self-awareness with an engineer's eyes. We are interested in artificial consciousness in the sense that the organization of a machine and its information flow yields a personal self-awareness, quite analogously to that in human beings. To this end, in the 1950s, right at the beginning of the computer age, "artificial intelligence" (AI) evolved. However, "out of a soup of ideas on how to build intelligent machines the disembodied and non-situated approach of problem-solving search systems emerged as dominant" (Brooks 1999, p. 146). Thus, countless and often very elaborate programs were written, trying to mimic man's extraordinary cognitive abilities. Yet when built into machines, i.e., when these programs served as control units of robots, they only succeeded in very restricted environments.

Because of the arguments outlined in the last sections, that does not seem to be a coincidence. Following this train of thought, it seems to be wise not to begin with soft- and hardware, but rather with a robot, the latter being much closer to the biological starting point that led to consciousness in the natural world. A robot is a machine of a special kind: It has a body equipped with sensory and motoric devices, and it is situated in a certain environment. If it is to act without the help of a foreign centre of control, it also needs an embodied "central processing unit" – a brain - that does the major part of the information processing necessary. Although these ideas are rather straightforward, it has taken decades for grounded cognition (Barsalou 2008) and mobile robotics to become more popular.

Typically, a robot is designed towards a goal. For example, if an animal wants to proliferate, it must survive for a certain amount of time and in order to do so it needs to gather food. Suppose its energy supply consists in a certain plant. Thus its task is to detect this plant and move its body toward it. Notice, that this "feeding behavior" will only be successful, if the sensory stimuli and the motor (re)actions are associated in the right way, i.e., if the program linking sensory input to behavioral output is able to guarantee sufficient food supply.

The program view

In total generality, one may think of a program that steers a body (O), i.e., that connects the input stimuli (S) with some behavioural response (R). There may be much processing going on in between, many stimuli might not lead to any action at all, and it could be very difficult to understand how an action came about or how stimuli, reactions and internal information processing are interrelated. Moreover, the sensomotoric program can be implicit, in particular hard-wired into the structure of a neural net, or it can be rather explicit, e.g., a long list of IF – THEN – statements, the IF-part containing all kinds of (external and internal) stimuli, the THEN-part containing all kinds of (external and internal) responses. However, whatever the details, the fundamental task and thus the basic solution are equivalent to both: natural and artificial systems. There is a certain body, situated in a particular environment that needs to be able to process information, turning relevant stimuli into adequate motoric output. Thus, in order to survive, it has to collect data, draw conclusions, and act appropriately.

Following the path of evolution, in primitive animals, there only seem to be a few fixed lines of code. That is, the sense organs are primitive, and the behavioural repertoire is very limited. As is the program connecting the two: certain stimuli will lead to foreseeable answers. Biologists call such a situation (behavioural) imprinting. Once an animal has adopted a certain behaviour it will display it over again and it is not able to change it. Searle's Chinese room is quite similar: A rather complex but fixed program receives questions in Chinese and answers them appropriately.

In other words, without further provisions, an animal or robot equipped with a constant program cannot learn. It is restricted to a possibly large, but fixed, way to perceive and to act upon its environment. Thus, every day, millions of flies die in flames since their program instructs them to approach bright lights, and countless generations of singing birds have been raising cuckoos, falling prey to the irresistible red colour of their parasite's throat.

Being able to learn is tantamount to saying that the sensor-motoric program possesses some *plasticity*. The program code – be it implicit or explicit - is not fixed (write-protected or closed) but open, subject to change. In the simplest case, a stimulus-response relation may be altered, i.e., an existing line of the program can (be) change(d). In particular, some stimulus may cause a different response than before, or a certain response may be the consequence of a new stimulus. More advanced cases would be existing lines being deleted, or new lines being added to the program.

Because the program resides in a body, located in an environment, there is also some *circularity*. Unlike a computer program, whose input and output are distinct, only connected by the program, the behaviour of the body may serve as the system's *input* in the next period of time. A response can become a stimulus, and the more often and the faster this happens, the more pronounced the feedback-loop. Weng (2009) points out that a system's output need not only consist in overt external behaviour, there could also be "internal effectors" working within the body. Moreover, external sensors can be supplemented with internal sensors, again strengthening feedback.

In total, there are a number of ways - chains of events - via which a program within a body may act on itself: It may, rather indirectly, affect its environment which later on has some consequences for itself (think of an echo in the mountains), it may, in a more direct manner, change the state of the body, which subsequently influences the mental state (think of alcoholic beverages). Yet the mind could also have a direct impact on internal affairs without the help of external feedback loops. For example, sombre thoughts may result in an emotional reaction, e.g., depression, whereas some meditation may evoke positive thoughts. Finally, the program may act directly on itself. Some output may, without other 'stations' being involved, directly act as its next input (electricians would call that a short circuit). Moreover, it may even happen that some line of the (open) program is actually writing the next line to be processed.

Multi-layered systems

Let us look at the program, connecting input and output, in more detail. In general, the information flows from the sensors to the effectors. Thus a fully developed program should consist of three distinguishable parts: a sensory part, dealing with the incoming information; a second part, analysing, integrating and drawing conclusions from the data; and a third part, responsible for the actions. In order to survive, the sensory part should not be a muddle of sensations, but should supply the individual with a well-ordered model of the world, within which the body is located. In particular, some representation of the body must also be an element of the world-model, and is of paramount importance: The program is located within the body, the information flows through the program, connecting sensory input and effective output, the body has to be steered successfully in the outside world, and it is the centre of all sensations and actions. This is what contemporary information science calls "embodiment", and the crucial program is called an "agent".

Again, a straightforward approach would be to devise an ingenious program that is able to cope with all kinds of situations, stimuli and responses. In addition, it should be able to learn efficiently. Thus the robot must be equipped with a lot of exquisite hardware (computational power) in order to sustain a very elaborate software system. At the very least, one needs:

a) an elaborate “world model”, integrating and augmenting sensory input
b) an impressive set of “thinking rules”, including many heuristics
c) a bunch of (flexible) behavioral strategies and tactics,
d) Meta-rules, enabling the system to change the way information is processed, and possibly the whole setup, if necessary.

Thus results a quite ambitious architecture with an emphasis on large, elaborate and specialized strata (cf. Brooks 1999: 4):

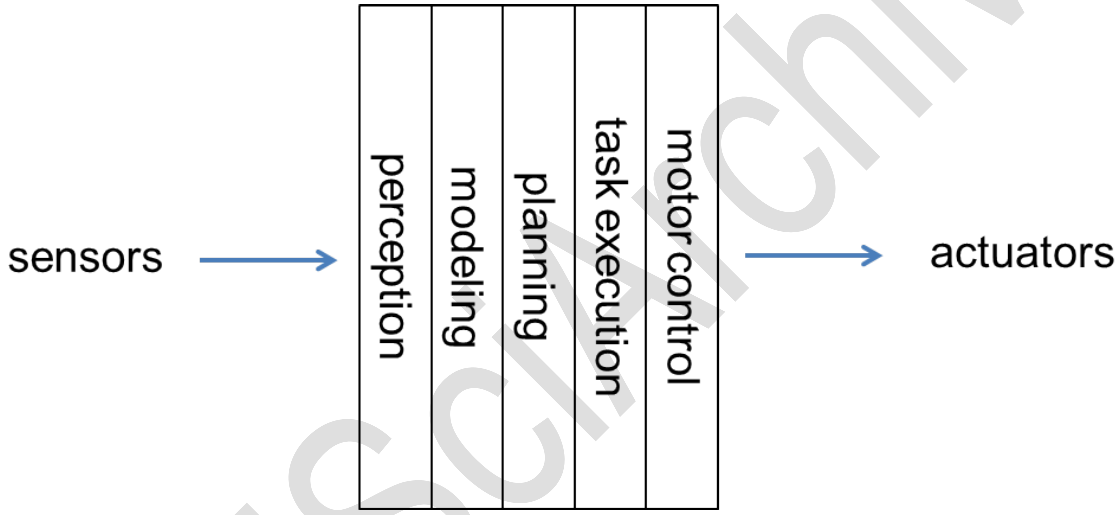

Illustration 11 (Vertical stratification: Perception – Thinking – Motor actions)

More or less explicitly, this paradigm was the hallmark of classical cognitive psychology and artificial intelligence. In countless efforts, these fields tried to understand the cognitive rules and setup used by animals in order to apply the results to robots. But again, decades of effort have not brought a real breakthrough, but rather disappointment on a large scale. Although robots have been equipped with the best hardware available and hundreds of thousands of lines of sophisticated code, their resulting skills have remained rather limited, if not miniscule. Despite various “IT revolutions” since the 1950s, and grandiose announcements, robots are still primitive.

Instead of trying to perfect this kind of “brute force attack”, in particular, building more and more complex “world models” and sets of cognitive rules, thereby increasing the computational burden, Brooks (1999: 5) changed the paradigm (the “cognitive landscape”). That is, his ingenious idea was to turn the task upside down:

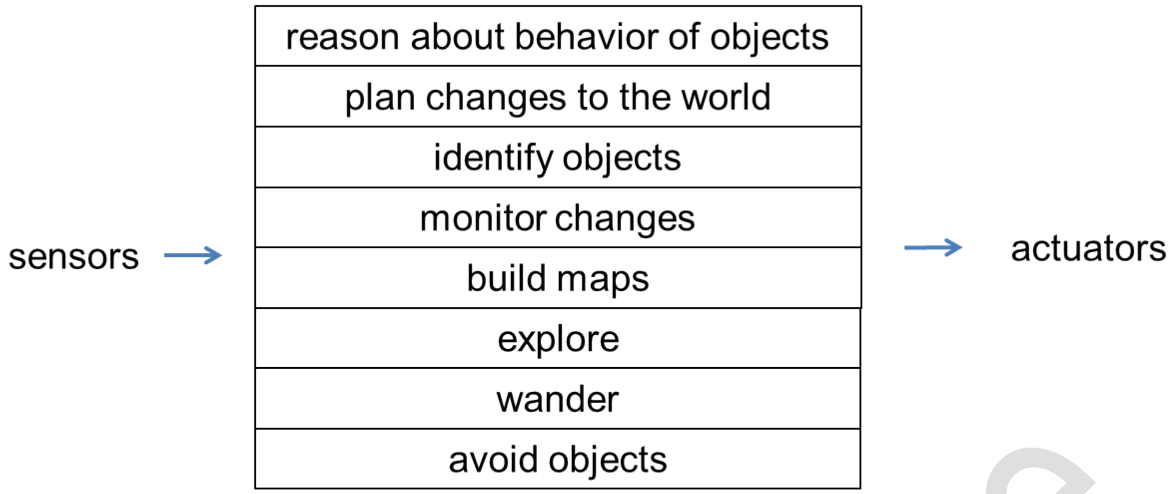

Illustration 12 (Horizontal tiers: Information processing on several levels)

Thus he could program simple robots in an efficient way. They were able to succeed in more complex environments than their predecessors and could react in due (i.e., real) time. Notice the general advantages of this kind of stacked architecture:

1. It is modular, i.e., a certain task can be broken down into small sub-problems and corresponding modules, specializing on specific jobs. Moreover, the modules and layers may, to a certain extent, act independently of one another. In particular, there is a lot of parallel computing.
2. It is hierarchical, with complexity increasing from bottom to top. Thus, rather primitive or standard behavior can be delegated to lower-level automatisms, yet higher modules, much more complex but drawing on the pre-processing of the lower tiers, may step in when appropriate
3. Given some input, there can always be a behavioral answer; i.e., fast responses – reflexes - at the bottom; emotions, encoding imprecise heuristics and vague strategies in-between; and cognitive responses, taking into account a lot of information in a rather explicit way, towards the top. At the very least, such a system does not stall when matters are urgent, and may find more suitable responses if time permits.

It seems to be no coincidence that this kind of elegant and flexible organization can indeed be found throughout natural and artificial systems. Multi-layered, modular architectures have become the most important IT-environments (in particular the n-tier application architecture, and the ISO/OSI-model of information interchange), but it is also well known that the nervous systems of animals are organized in a similar way:

"It is a basic principle of neuroscience that the cerebral cortex is divided into segregated areas with distinct neuronal populations…This *anatomical* classification of neural areas can serve as a basis for classifying cortical regions according to their function" (Bermúdez, 2014, p. 248, emphasis in the original). These modules (Fodor

1983) are readily combined into an overall hierarchical structure (see below, in particular table 13, but also Marr (1982)). Carrying the analogy further, in the natural as well as in the technical world, the basic building blocks are quite simple (neurons, logical circuits), operating systems keep the body alive and on track (homeostasis), languages directed towards problem solving are important, and despite enormous complexity, common architectural principles keep chaos at bay. Finally, it almost goes without saying that software has to be supported by adequate hardware. Since psychology builds on physiology and anatomy, our species should and indeed does possess the highest encephalization quotient.

Learning

There is yet another major advantage of a multi-tier architecture: It can be extended gradually, i.e., on the top of the existing edifice, further layers may be added one by one. This secures the overall functionality and stability of the system, but also opens up qualitatively better ways of dealing with the environment. Thus more and more sophisticated layers have evolved, with abstract, clear and rational thought being the pinnacle.

The general developmental direction is bottom up: from primitive to complex, and from implicit to explicit. Thus, in order to understand and develop such a system, it seems to be a good idea to start from scratch, i.e., with the very first layer, with simple responses, and rather primitive tasks. Straightforward observations corroborate that babies, but also successfully developing robots, begin with the basics (e.g., gazing, grasping, sitting), and spend much time acquiring fundamental skills.

In this endeavor, prior information built right into the design of the system helps a lot. Haykin (2009), pp. 58-59, gives a number of reasons why: A specialized neural network is smaller than its fully-connected counterpart, needs smaller data sets for training, learns faster, and the network throughput (information transmission) is accelerated. In the most extreme case, an incarnated structure or feature need not be learned at all, and since a built-in structure is cheapest, it can also be expected that nature applies this trick whenever possible.

Still, much has to be learned in the concrete environment where the system is living. Since a lot of energy and effort have to be spent in order to get a fully-functional result, rather typically, mental “growth” is tedious with new features emerging gradually. Noticing that learning by imitation is easier than learning by reflection which, in turn, is less dangerous than learning by bitter experience (cf. Confucius), it is no surprise that copying successful strategies applied by others, e.g., imitating what they are doing, is the most common form of making progress. Thus we are constantly checking the activities of others, their mental life included (thus creating a “theory of mind”, i.e., we are able to detect and understand expressions, intentions, beliefs of others). Even on the hardware layer, we are equipped with so-called mirror neurons that have specialized in detecting interesting behaviour in others. Recently, robotics has begun to implement these devices and ideas in machines (cf. Cheng 2015, Chapter III).

More generally speaking, Brook's layered architecture and his main insights were the springboard for "developmental robotics" (Weng et al. 2001, Asada et al. 2001, Asada et al. 2009): a promising field which has become more and more popular. Conceiving "a roadmap for cognitive development in humanoid robots" (Vernon 2010), researchers in the field do not try to teach their robots particular tricks for a restricted environment. Rather, their emphasis is on curious "playful machines" (Der and Martius 2011) actively exploring their environments. Given intrinsic motivation (drive(s)), a robot moves around, gathers experiences, constantly learns and develops step by step along its own path.

In this process, prior structure (talent) is basic, imitation and repetition help a lot, and personal experiences including typical mistakes are inevitable (you cannot ride if you have never fallen off your horse). In addition, however, much learning is explicit, and as a consequence, thinking and understanding (at least in humans) should not be underestimated. Learning by reflection may be noblest, but much more importantly, one has to think by oneself in order to become smart. To grow strong, the brain (like muscles), needs permanent, well-targeted training. As we all know from the education of children, this kind of learning can be accelerated by suitable hints, specific rewards and (at times) punishments from a more experienced person, i.e., a teacher.

All in all, "guided self-organization" (Der and Martius 2011, Chs. 12-14) seems to be the most promising method of development. Along this way, children also detect and cultivate internal processes such as attention, motivation, estimating and weighting, helping them a lot in acquiring explicit knowledge about the world they live in. With knowledge comes reasoning (Mareschal et al. 2009), but also deeper insight and evaluation. Finally, having understood many explicit features of the natural and the social worlds, and having matured on the inside, they have/are grown up, able to deal with challenges of all kinds in a sophisticated way – their way.

"Self-inflating" systems

There is a common general idea in pedagogy as well as in developmental robotics, which is to create a "self-developing" system, i.e., a system that remains stable although it is able to learn immensely (Cangelosi and Schlesinger 2015). That's quite a challenge, since altering a complex system straightforwardly threatens its functionality (never change a running system…); yet it is crucial to add, run in and restructure features, modules and layers. It is a formidable task to move from primitive to elaborate and complex, from reflexes and blind imitation to critical reflection, without shutting down the system even once.

At this stage, it could be a good idea for computer scientists to study the theories of developmental psychology. Of course, psychology uses a different jargon, and, owing to its subject matter, it is quite often rather vague. Nevertheless, hierarchical conceptions are a commonplace in psychology (e.g., Freud's psychodynamics with the id, the ego, and the super-ego, Maslow's hierarchy of needs, Kohlberg's moral tiers, Jensen's theory of intelligence, and Loevinger's ego development, to name but a few.) Moreover, developmental theories capture important features of successful and pathological mental development. Adopting crucial principles of human (and animal) development may help find strategies to further the mental development of artificial information processing systems, in particular robots. Let us look at two theories in more detail.

Kurt Lewin (1939) proposed the first "field theory", that is, he located the individual within its so-called "life-space", i.e., a complex environment endowed with social "force fields" of all kinds. Within these boundary conditions an agent acts and develops. Probing its environment, it learns and moves from the primitive and even rudimentary to the sophisticated and complex. For example, first a child learns to lift its chin and its chest. Then, it is able to sit. Subsequently it crawls, stands, and finally walks. Most of these abilities have to be practised many times, and are first accomplished with and then without the help of others. Quite similarly, Der and Martius (2011), p. XI, write: "…behavior generation in complex robotic objects is improved and stabilized by taking brain, body, and environment as a whole. The playful unfolding of behavioral patterns offers a new way of getting the embodiment of the agent involved."

Jean Piaget's "developmental stage theory" describes in some detail how children evolve cognitively. Altogether, there are four qualitatively-different levels, and a more basic level has to be "completed" in order to move on.

Small children (up to 2 years) start in the "sensorimotor stage" when they first gain experiences with their sense organs, their movements, and the co-ordination of the two areas. In other words, babies and toddlers develop basic sensory and motoric abilities, and learn to close the sensorimotor loop (e.g., object perception, face recognition, gaze following, selective attention, manipulation, and locomotion). Finally, they learn to "internalize" concrete actions, i.e., instead of handling concrete objects they start to operate with mental tokens - "doing" is partially replaced by "thinking". Mental objects may still be present when sensory attention shifts.

At the age of 2-7 years, in the "pre-operational stage", children are egocentric in the sense that they know about their own position in the world. However, it takes (them) a long time to transcend this particular perspective, i.e., to learn that there are different perspectives (e.g., those of others) and to mentally assume the other's "point of view". Attention is also restricted to a single feature of a situation (I am hungry, and I want to eat now). Moreover, the thinking of small children is still magical and animistic. On the one hand, objects are endowed with human capacities (e.g., a “bad table” wants to hurt children). On the other hand, powers beyond their reach voluntarily influence their lives (e.g., Father Christmas bringing toys).

Between 7 and 12 years, children reach the “concrete operational stage”. That is, the number of mental objects considered at a certain moment grows, they become able to simultaneously grasp various aspects of a problem, i.e., they get an overview, and may thus solve increasingly complex problems in concrete situations.

At the age of 12 to 15, they finally reach the last, the abstract, level, the “formal operational stage”. Juveniles learn to work with arbitrary assumptions (“what if…”), and are able to check the consequences of hypotheses systematically. Deductive and inductive reasoning become possible. Finally, they may leave familiar realms, and consider instead general and abstract problems, quite removed from their daily experience. “Coming of age” means to receive full authority of one’s own life and to be in charge of further self-development.

Development via interaction

How are these remarkable goals achieved? The most basic point in development is the permanent and intense interaction of the individual - and its programming - with the environment (Pfeifer and Scheier 1999). It is this ceaseless, and at the same time rather selective, interaction between the individual and its environment that pushes development forward. Focusing on either nature or nurture, or trying to separate their impact, seems to go astray or even miss the target. In other words: Because of situatedness, the question as to how much is inherited or due to the environment is rather misleading. Talent may be important, but without appropriate training, it will never flourish; we need to understand the interaction of the organism with its environment in order to achieve optimum results.

Since learning is “online”, **curiosity** - resulting in the (systematic) exploration of one’s environment - is very important if not crucial. It pays to learn much about the world we are inhabiting, to explore and understand it. Although this kind of active behaviour is more dangerous than sitting idle (the so-called “lazy robot effect”) and may at times kill a “playful machine” or the proverbial cat, at the end of the day, intrinsically motivated robots have learned more about their environment, and are thus more successful than sluggish ones (cf. Der and Martius 2011, Cangelosi and Schlesinger 2015, in particular Chap. 3).

In more detail, according to Piaget, a “schema” is a basic building block of human knowledge, both in the cognitive as well as in the behavioural arena. In Piaget (1952) he defined a “schema” as “a cohesive, repeatable action sequence possessing component actions that are tightly interconnected and governed by a core meaning.”

Moreover, there is a simple mechanism which could be called the “assimilation-accommodation *loop*”, extending and sophisticating schemas:

1. Internally, one starts out with some rather primitive schema, on the most basic level it is an inborne capacity (e.g., the ability to grab or to make noise)
2. *Assimilation* refers to the process of fitting new information into pre-existing cognitive schemas. That is, some new experience is processed with the help of the existent framework, adding a new chunk of information to the latter, but not altering the received schemas.
3. However, at times, it may be necessary to change existing cognitive schemas. This process is called *accommodation*, and happens when the existing set of schemas does not work with a new object or situation. In particular, a crude schema (idea, concept) may thus become more differentiated.
4. With every loop of this kind, the (internal) cognitive system becomes more complex. Thus the agent is able to deal with more objects and situations, or to treat a particular problem in an increasingly sophisticated manner.

Vygotsky (1978) described this process of “extension” in the area of assisted learning: He distinguishes between a zone of independent performance (i.e., what the child can do alone), a zone of assisted performance (i.e., what the child can do with support), and the zone of “proximal development” in between. Therefore he advises that efficient teaching be directed at the latter. In other words, it is like “scaffolding”, supporting the development of attainable skills, finally leading to autonomous mastery.

A progressive mental system depends on major feedback loops connecting it with its environment, but also internal loops grounding it in more basic units, including the physical body. Its growth is not random, but well-structured: Auto-didactically, but also assisted by teachers, a multitude of modules dealing with particular tasks are developed sequentially. In order to become fully functional, they have to be linked and integrated into a large network, finally leading to comprehensive mental layers that rest on each other.

Psychologists also noted that creating such a sophisticated, flexible and, at the same time, stable mental system isn’t an easy task (e.g., see Rubenstein and Rakic 2013). In particular, there needs to be an equilibrium between internal processes and the incorporation of (new) sensory input. Behavioural responses must be calculated in time, etc. Finally, all these organizational structures and processes, heavily dependent and building on each other, and interwoven by countless feedback loops have to be precisely parametrized in order to avoid over- and under-stimulation. This seems to be the reason why some authors refer to the fitting image of a “fine-tuned orchestra” to describe the immense amount of precise feedback, control and timing necessary.

Not surprisingly, for their mental abilities to develop and flourish, animals with large brains need much time, and even adult teachers to guide them. Homo sapiens has reached the far end of this evolution: Our children require several decades of education to finally reach cognitive and personal maturity. On the one hand, this long timespan seems to be due to the sheer number of abilities that have to be acquired and mental structures that have to be erected. On the other hand, there is a simple physical reason: Much of the available energy is consumed by our nervous system. Thus, when the brains of children are developing fastest, i.e., when they are about five years old, physical growth is slowing down (Kuzawa et al. 2014), and has to be postponed until adolescence. In this sense, man is a very powerful mobile IT-system whose development touches upon physical limits.

The following table summarizes the mental edifice:

Mental Layer	**Processes**	**Building blocks**	**Main location**	**Evolutionary stage**
Apex	Structured reasoning, language	Concepts	Associative areas of the cerebrum	Homo sapiens
Top	Cognitive	Cognitions of any kind, in particular, images	Cerebrum	Primates, some whales and birds
Intermediate	Emotional	Emotions, e.g., fear	Limbic system	Mammals
Basic	Reflexes	Drives, rigid procedures	Brain stem	Reptiles
Elementary	Fundamental Responses, often 0-1	Communication pathways between neurons	Neurons, and sets of neurons	Animals with nervous systems

Table 13 (Mental organization of biological species / natural systems)

Towards consciousness: Self-representation(s)

Given a rather complex and multi-layered system, it remains to explain how self-awareness fits into this picture. In more detail: How does such a system beget an idea of its own existence?

First, it is crucial that the individual is a part of the whole situation. In particular, its mental equipment is a crucial part of the information flows running within the body but also between the subject and its environment. In other words, the layers of the "central processing unit" communicate with one another, and, on each level, there are strong feedback loops between each layer and the environment.

Second, since the individual is a part of the whole system, there exists a token for it on every level. Depending on the level, this token is an emotion, an image (or a more complete sensory impression, including, for example, a voice) or a concept (i.e., a word for oneself). Since the individual being is crucial and since it is located in the centre of action and perception, these tokens – placeholders - are out on a limb.

Third, combining the properties just described, it is straightforward to conclude that the individual "notices" itself in the sense that the various tokens detect themselves due to tight circularity. A certain token, being a representation of oneself (or one's self, respectively) on a certain layer, looks ahead and therefore perceives the back of itself. Matisse's famous picture "not to be reproduced", painted in 1937, exemplifies this situation perfectly: A man is looking in a mirror. But instead of seeing his face, he is looking at his flipside.

This organization of the mind seems to fit quite well to our impression that higher developed animals have some insight into what is going on. They are not just automata, following pre-adjusted instincts. Rather, they are flexible, able to adapt to new challenges and environmental conditions and, in the sense just described, seem to have an idea of themselves. In other words, they are thus able to distinguish (to a certain extent, more or less precisely) between themselves and the rest of the world.

Now it should come as no surprise, that the most elaborate theory of self-awareness, developed by Damasio (1995, 1999, 2010) in a number of books, is hierarchical, in more detail, it is a three-layered theory of consciousness:[1]

1. First, there is the protoself. All sensory inputs are integrated in one basic brain structure: the hypocampus. Its first task is to keep the system alive, i.e., to guarantee homeostasis and homeokinesis.
2. Second, in higher developed animals, feelings emerge. An emotion is an unconscious reaction to any internal or external stimulus which activates neural patterns in the brain. A 'feeling' emerges as a still unconscious state which simply senses the changes affecting the protoself due to the emotional state. Core consciousness is the feeling of knowing a feeling. (Note the reference of an emotion to other emotions. Thus emerges a second layer: Feelings based on other feelings.)
3. This is also where Damasio puts his third layer of self-reflectiveness which he calls "extended consciousness". Although he thinks that language is not essential in its constitution, he emphasises the role of memory in order to build a particular personal perspective, including ownership of thoughts and the power to manipulate mental items. Thus, we understand in a much more explicit way than before who we are (i.e., our properties), our position in space (in nature and society), and the arrow of time. In particular, there is a stable personality with an autobiography.

Although the author of this contribution agrees with most of Damasio's conception, I think that language is crucial. The main reason being that due to circularity and thus

[1] For a closely related account see Donald (2002).

self-reference, it is rather straightforward to perceive one's flipside. However, a really strong "mirror" is needed in order to look oneself straight in the eye.

In the first place, nothing spectacular happens when some sensory input is associated with a certain sequence of sounds. It is a long way from isolated sensory input to a continuous, multi-facetted impression of the world surrounding some body, and in analogy to the world model before, it is also a long way from syllables and single words to a well-structured language, able to describe almost everything in a concise way.

Our closest evolutionary neighbours – apes – are hardly able to learn more than a few dozen words, and they are very slow in using them. One has to produce and utilize language in an advanced manner, in order to take the next crucial step (already described in some detail in the first and second section).

Full Self-Awareness

As far as we know, newborn babies are hardly aware of themselves. Thus, the decisive step occurs later and can be observed. When small children label all the objects in their vicinity, they also use a name for their own body. Yet they need not understand that this concept is special. However, when the world model and their language proficiency improve, they have two efficient ways to describe matters (helped, of course, by the more basic layers and their environment).

In both, the sensory and the verbal realms, the representative of the body is a well-defined, crucial entity: its image and its name are located at the centre of operations; all sensations, actions, and mental processes revolve around these tokens. Ordinarily, the formation of a concept is like the development of a dictionary – one simply learns which word has to be used for a certain sensory input. But since the program is continuously observing the body in which it is located, there is no distance possible. Moreover, combining the particular body image with the corresponding, equally remarkable, word yields a truly exceptional concept; a concept representing "the centre of the world" (more precisely, the world as seen from my point of view).

Upon changing perspective, typically the name of an object remains fixed (an apple remains an apple, no matter who is talking about it). However, when others talk about themselves, they also always use the same word – "I", and a number of different words for others. Adopting this practice distinguishes "me" from all the other objects and phenomena around. (In other words: observing others, how they act and how they perceive themselves may serve as a "role model" or as a catalyst in detecting one's own self or quite simply "oneself".)

Thus "I" is not really "my" name, but rather a particular point of view that should be distinguished from others, be they objects or persons. Being inseparably connected with private feelings, observations and actions, it stands for my (complete) subjectivity. This and only this concept represents myself. Full self-awareness means possessing such an extraordinary concept and understanding its entire meaning: First, “I” stands for my humble self, and not anybody else(‘s). Second, my self is not void but inseparably linked with exclusive, body-related information. Third, “I” is / am defined by a distinct, stable line separating myself from the rest of the world.

Our line or argument and common knowledge may be summarized in the following table:

Mental Layer	**Crucial Processes**	**Mental token for oneself**	**Personal insight**
Apex	Language	Precise concept: I, combined with a stable division between one’s self and anything else	Full personal awareness and situatedness
Top	Cognitive	A complete “image” of the body’s current state	Extended consciousness
Intermediate	Emotional	The feeling of what happens: Integrated multisensory maps	Core consciousness
Basic	Reflexes and drives	Taking notice of one’s existence	Protoself
Elementary	Fundamental responses	Almost non-existent	None

Table 14 (Levels of consciousness reached by biological species)

Reading the last table horizontally reveals that the multi-layered architecture yields qualitatively different descriptions of the world. Therefore, due to situatedness and thus circularity, different representations of oneself emerge. Depending on the level, a more or less clear understanding of one’s self is reached, i.e., a certain kind of self-awareness follows suit.

Reading the last table vertically shows several tiers, resting on one another. Evolutionary speaking, each new layer is associated with a characteristic kind of process, language being the last and thus uppermost. The descriptions of oneself available (one “self” on each level) become more and more explicit and precise. They build on each other, until finally, conceptually clear, stable and comprehensive personal awareness is reached.

Note that the whole edifice as well as personal awareness is differentiated and integrated. On the one hand it makes sense to distinguish several kinds of consciousness, on the other they are all combined into a single mental unit. We will elaborate this important point in the next section and consider the inverse of development, i.e., degeneration, first.

Degeneration

An agent's organization, be it self-conscious or not, is based on the body being located in some environment with information flowing constantly from the sensory devices via internal links to the motoric output. Adding to this forward stream a strong feedback loop (from the (re-)actions taken to the consequences perceived) makes information circle. It is no coincidence that for homo sapiens, clear conceptual thinking, combined with tight "eye-hand coordination" (i.e., eye-guided hand movement and hand-guided saccades of the eye) proved to be crucial for his evolutionary success.

When this basic process is weakened or breaks down we should not be surprised that self-awareness is affected too. Why should we be aware of ourselves when we fall asleep? When sleeping, sensory input is almost irrelevant, and information processing in the brain may take completely different paths. Instead of a multitude of actions being the result of internal processing and ceaseless external inputs, it rather seems necessary to quit the run-in paths followed in conscious life and let information flow in a less entrenched manner, which could be the reason why we are particularly creative when asleep. The same happens if we cut the input stream of information as happens in meditation and sensory deprivation experiments. Backpropagation, i.e., a (partially) reversed flow of information could also be important, enabling output-driven re-programming of the neural nets connecting the perceptual layer with the desired behavioural result.

Of course, self-consciousness also fades away if the neural basis is damaged, e.g., under the influence of drugs, during an epileptic episode, or because of degenerate neural diseases. Notice, however, that due to the overall hierarchical structure, a breakdown typically does not destroy all layers at once. Much more frequently, the erosion of the mental edifice begins on the top level and gradually proceeds to the basics. First, language may simplify, orientation may deteriorate, or complex manual skills may fade. Later, situatedness may get lost, family and friends may no longer be recognized, and the personal identity weakens. Later still, the cognitive level may be gone completely, and when the emotional layer disintegrates, all that is left are rather primitive basic drives and reflexes (e.g., thirst and hunger). In the worst case, all that is finally left is a bodily shell without mental life.

In a life-span perspective, the "office of schemas" has humble beginnings, evolves immensely, reaches full complexity in adulthood, withers and finally fades away. Mathematically speaking, mental degeneration is the inverse function of cognitive development: first we grow up, then grow old (and finally pass away). Thus the mental realm of a not yet fully-developed child may resemble in much detail an elderly person's eroding cognitive landscape, transforming the rather superficial observation that "old people become children again" (Hamlet, 2nd scene) into a deep Shakespearian truth.

4) Complex systems

More is different (Anderson 1972)

There is yet another, more principled and abstract line of argument leading to consciousness. Apart from putting this remarkable phenomenon in a larger perspective, it gives some concrete ideas about how it develops and how it is structured. Applying these insights may help to program “artificial intelligence”.

Emergence of new properties

For a long time, reductionism ruled. That is, in order to understand some phenomenon, it is crucial to break the phenomenon down into its constituents and analyse their causal relations. Having thus grasped the inner workings of a mechanism and its main elements, one should at least be able to predict its major results. In a sense, this is the overall “modus operandi” of science: analyse an interesting phenomenon in detail, until you have understood what is going on.

However, “the ability to reduce everything to simple fundamental laws does not imply the ability to start from those laws and reconstruct the universe. In fact, the more the elementary particle physicists tell us about the nature of fundamental laws, the less relevance they seem to have to the very real problems of the rest of science, much less to those of society. The constructivist hypothesis breaks down when confronted with the twin difficulties of scale and complexity” (Anderson 1972, p. 393). In other words: having understood the details typically does not imply the big picture. Why?

The reason is that “at each level of complexity entirely new properties appear, and the understanding of the new behaviors requires research which I think is as fundamental in its nature as any other… each level can require a whole new conceptual structure. Psychology is not applied biology, nor is biology applied chemistry… the whole becomes not only more than but very different from the sum of its parts” (Anderson 1972, pp. 393-396).

In the last twenty years or so, the associated philosophical position of *emergentism* has gained ground (for a short introduction and a long list of references see de Souza Vieira and El-Hani, 2008), opposing reductionism, and stressing the importance of organization and supervenience. A particularly interesting account of emergent phenomena is given by Deacon (2007). The most important ideas, however, originated in the natural sciences. For an overview see Èrdi (2008) but also Murphy et al. (2007).

Given this viewpoint, one should not expect that self-awareness may be explained by way of reducing it to some fundamental physical law like Heisenberg’s uncertainty principle. On the one hand most scientists would agree that the brain can be reduced to standard physical particles and forces, and that neurons are the basic elements to be considered. (There is no particular mental “stuff”, or “vis vitalis”, etc.). However, on the other hand, there is also a consensus that the brain’s organization - its anatomy and physiology, i.e., its structure and dynamics - is crucial. That is, despite material reduction, it is a very persuasive idea that self-awareness is an emergent property on a certain level of (biological) evolution, more precisely, of a certain kind of (complex) organization.

One of the main theses of this contribution is the claim that the systematic use of a rich, natural language leads to completely new features, in particular self-awareness. More specifically, the above arguments suggest that our personal self is the consequence of conceptually-clear, circular information processing, inclosing a preeminent mental token for the person processing the information. In a nutshell, the phenomenon of self-awareness is a major consequence of a sophisticated mental organization, i.e., the multi-hierarchical and self-reflexive flow of information in situated robots, based on precise chunks of information about the agent and its environment.

Building the final layer

The crucial problem for nature and thus also for engineers and computer scientists consists in constructing a stable hierarchic structure, governing the dynamic flow of information (within the robot, but also in relation to the world outside). Piaget’s idea of assimilation and accommodation describes an elementary, recursive loop, extending the mental system. However, it does not explain how new modules or even layers are created, developed and integrated.

With the help of the above reasoning, one may explain how the crucial new feature of language was added, i.e., how the language sub-system may have developed,[2] and how this innovation led to the unique human mind:

a. **Reaching the tier of language.** Great apes possess a sophisticated visual image of the world. Of course they are able to hear and to listen, and can produce a broad range of sounds. Linking a specific sound to a particular sensory impression creates a primitive concept. Thus, above the familiar sensory model of the world they inhabit, a new level is reached. First, perhaps, accidentally, but, if sounds with a particular meaning, words and concepts bring about an evolutionary advantage, it pays to repeat the process of concept-formation.
b. **Consolidating the new tier.** Doing this a dozen or a hundred times produces a vocabulary which is enriched by every new concept created. Combining the words according to rather constant rules produces an even more powerful way to describe persons, phenomena, and habitat. Thus, a new tier emerges. On

[2] Similarly, for every more basic tier, one may describe how the next layer was possibly built on top of the existing structure.

the one hand it is thoroughly grounded in bodily experiences (sensory perceptions of all kinds, but also motor actions), on the other hand, it has rules of its own (a grammar). Playing around with concepts and grammar creates the layer of language, i.e., a versatile tool with a complex structure and a "life of its own", able to describe and explain what is going on.

c. **Downward consequences.** The tier of language, built on top of the sensor layer, has a major backlash on the received structure. Numerous authors (e.g., Sellars 1970, Campbell 1974, van Gulick 1995, Deacon 1997 & 2007, Murphy et al. 2007) have considered it. In general, the feedback of a new emergent feature on the elements on which it is founded is called a 2nd order constraint. Since such "top-down" constraints can be very powerful, Haken (2006) uses the term "enslavement," Bunge (2003) speaks of "submergence", and Sperry (1973) notes that a whole may "overpower" its components. However, with respect to mental processes the most frequently-used term seems to be "downward determination" or "top-down causation" (or a combination of these words). In effect, the feedback leads to a complete
d. **Mental reorganization.** Previously, mostly in Sections 1 and 2, we discussed in detail how the closed feedback loop acts like a landslide, altering the mental landscape dramatically. Therefore a succinct summary should suffice here: First, due to circularity, there is a particular token standing for one's self on every tier (see Table 14). Second, in particular, there is a word for myself on the top layer, and an elevated image of my body in the visual domain. Third, due to the tight feedback loop between these layers, they act as "internal mirrors". Fourth, the self-concept and the self-image may combine, forming a stable and well-defined (personal) identity that is enriched by contributions of more basic layers (e.g., the protoself). Fifth, drawing a razor-sharp line between ourselves and the rest of the world, we become aware of our precise position in space and time.

The new organization which has thus evolved features a personal self, i.e., a clear sense of self-awareness, a definite idea of oneself, on top of the restructured mental edifice. This self is grounded in and based on several layers with their particular components (e.g., concepts) and internal structures (e.g., grammar). On the one hand - bottom up - the self incorporates facets of each of the more basic layers (in particular, it has a cognitive and emotional flavour), on the other hand it is an agent, controlling - top down - parts of them. For example, it is able to act actively (voluntary motor function), talk (conscious command of language), think (intentional use of cognitions), and direct its sensory alignment (focused attention).

However, each layer also has a "life of its own," and the farther away it is from the top, the more so. Owing to the organizational structure, we are able to cope with language best; words are right "on the tips of our tongues", ready-to-use. It is more difficult to control memory and general cognitions: We may not be able to retrieve a certain memory, rotating an object with the inner eye is tedious, and it is very difficult *not* to think of a pink elephant if told so. The motoric realm is divided into voluntary and non-voluntary motor functions. Since we have no direct access to the emotional tier, we are also not able to control emotions directly. If a person is sad, it does not really help to be told to cheer up, and if frozen in shock, it takes much voluntary effort to overcome paralysis.

Altogether, the edifice looks like a pyramid, governed by a personal self:

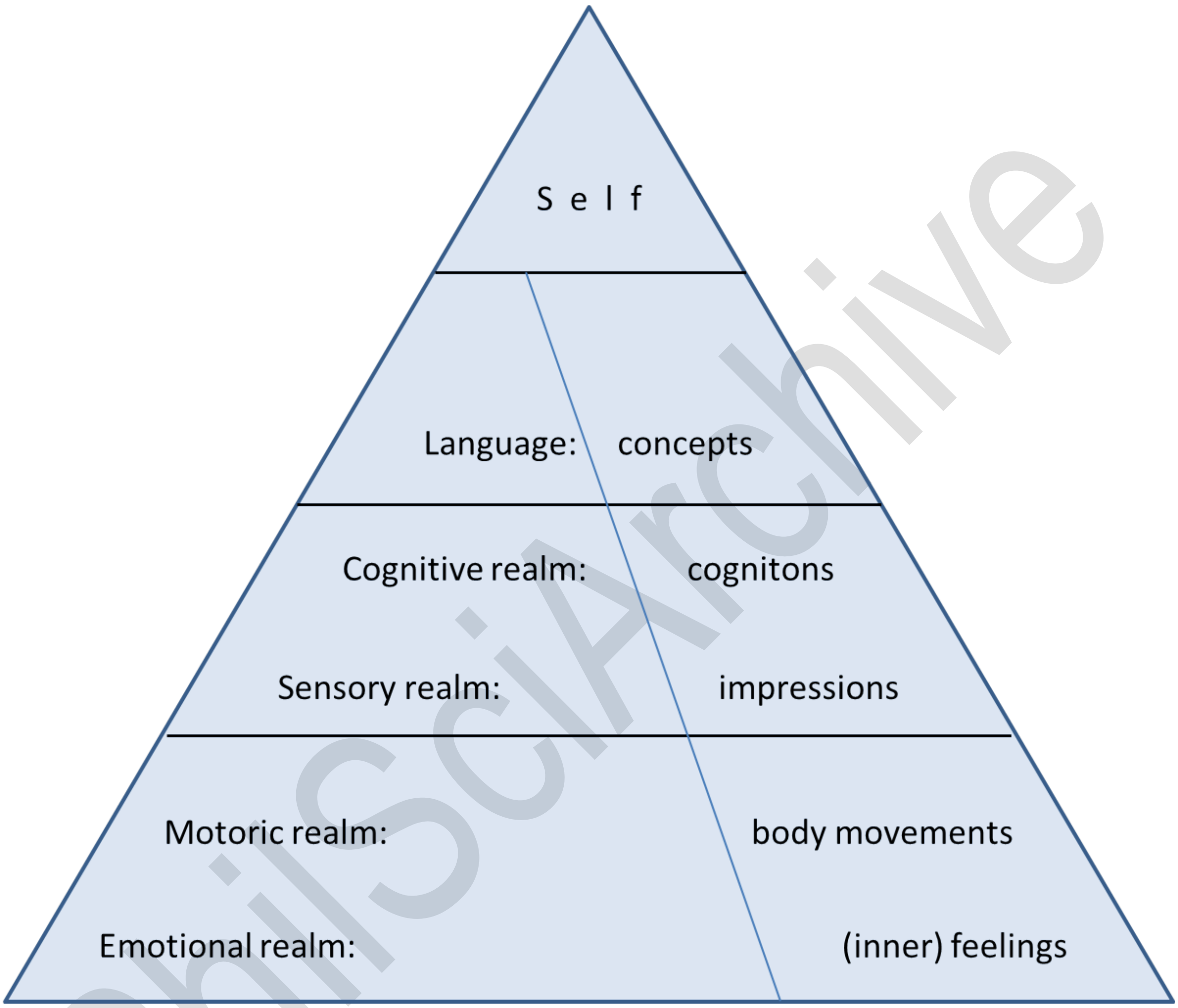

Illustration 15 (The mental pyramid: a multi-layered edifice with the self on top)

Further implications

It may be added that quite obviously, the faculty of language, combined with a sense of self, also greatly facilitates and improves communication with others. Thinking in tiers, straightforwardly, a social tier may be added on top of the above pyramid. In other words, due to language the link between several individuals of the same species is strengthened and more durable social structures than ever before can be built on conscious individuals. Thus man became the most eusocial animal ever. With the groups growing in numbers and stabilizing, this layer brought about sedentariness, systematic farming, division of labour, and a multitude of other historical traditions; in the end producing large societies, sophisticated culture and civilization.

Historically, one may distinguish three major moves: Hundreds of thousands of years ago, spoken language, leading to self-awareness, singled out the species of homo sapiens. Thousands of years ago, writing greatly increased our ability to store and pass on information, which founded civilization. Hundreds of years ago, the formal language of mathematics brought about science and technology, i.e., a much deeper understanding and command of all kinds of phenomena, which characterizes the modern era. Taken with a pinch of salt, language–based innovations (printing and the internet included) have exponentiated our ability to learn about the world and ourselves. We truly have become the symbolic species (Deacon 1997).

5) Free will

Who is in charge here? And who is doing all the work? (van Gulick 1995)

The theory developed above, in particular the last illustration, corresponds nicely with our self-evident “naïve” personal everyday experience. It fits with the modern idea of an autonomous agent, but also with the time-honoured view of “free will” (liber arbitirium as it has been called at least since the Middle Ages) that may be traced back as far as Aristotle’s “De anima” and Plato’s dialogue “Phaedrus” where they depicted the “I” as the charioteer of the soul. Contemporary psychologist Roth (2003) characterizes this idea by saying that the self is in superior command of thinking, planning and action, being a central decision and executive system in the mental realm. He also highlights self-monitoring, self-government and autonomy.

Although the concept of personal freedom has an air of arbitrariness and non-determinism, in particular in the philosophical debate (Kane 2011), it is mostly used in the above sense by natural scientists (Baumeister 2010). So the main connotation of freedom is the self being in charge, having a meaningful choice, being the owner of the mental edifice and the captain of the body. The contemporary challenge to the received top-down conception is bottom-up determinism. The more we have learned about the brain, the more it has become obvious that physical processes are the basis of the mental: All thought and emotion, perception and action, memory as well as personality, depend on anatomical structures and physiological mechanisms. For example, an extraordinary memory (fast, large and reliable) is needed to support language, with the hippocampus (Marr 1971) as well as the classic speech areas of Brocca and Wernicke playing major parts in this narration. In general, brain damage readily implies mental limitations. Thus it is straightforward to conclude that psychology is an epiphenomenon of neurobiology, and the striking results of Libet (1985) and others (in particular, Kornhuber and Deecke 1965) have been interpreted in this way. Given the question “do we consciously cause our actions or do they happen to us?” (Wegner 2002), many natural scientists exploring the brain “bottom up” opt for the second view. For a thorough discussion see Baer et al. (2008).

The synergetic computer

However, considering a common computer, it is obvious that problem-oriented programs near the top level drive the physical actions. In the end, the “locus of control” is the user, who - via the computer’s graphical user interface - tells the software and the underlying hardware what to do. The crucial flow of information consists in commands issued “top down”.

How can we explain such a kind of “free will” in self-referential, dynamical systems that we have studied? Very close to our account is the idea of a “synergetic computer” (cf. Haken 2004, Haken and Schiepek 2010) which consists of at least two layers, ceaselessly influencing one another.

The first crucial idea of synergetics is that the elements that make up the bottom layer may interact in a particular way, producing an overall pattern (which due to its regularity can be described by a few order parameters), forming the upper layer. The paradigm example is light: Unlike ordinary sources of light, a laser does not emit uncorrelated light waves. Instead, it produces a highly coherent single light wave. That’s the bottom up impact, i.e., the elements’ spontaneous self-organization into a larger and simple structure.

The second crucial idea of synergetics concerns the top-down impact, i.e., the consequences of the large structure (often represented by its order parameters) on the elements in the bottom layer. In the case of the laser, the coherent light wave forces single photons to oscillate in the same way. That is, the elements are no longer “free to do what they like”. Instead, they lose many, if not most, degrees of freedom and have to comply with the overall organizational structure (therefore the term “enslavement” mentioned before).

With respect to information flows in the brain, this account captures many important aspects. It is both elegant and there is much experimental evidence in favour of it:

1. We have stressed the importance of feedback, i.e., of an account that is dynamic as well as circular. Synergetics explains how a hierarchical system endowed with a feedback loop/ circular causality may emerge spontaneously via “self-organization”.
2. Looking at the lower tier, there are indeed coherent waves when areas of the brain work together (Singer 2007) which are to be expected when parts – via a common order structure – co-operate (automatic “consensus-building” in the words of Haken and Schiepek 2010). More generally, in this view the “binding problem” (how can different brain parts work together when necessary) is

solved via spontaneous synchronization bottom up, in particular frequency locking (Haken 2002, 2008).

3. Looking at the upper tier, there is an enormous degree of information compression, since a few order parameters suffice to describe the overall behaviour. It is well known that we do not store all details of a story or picture. Rather, we retain the most interesting, striking and characteristic features.
4. Since patterns may act as the building blocks for further layers, picturing a multi-tier system is straightforward.
5. Information processing within this system is both massively parallel (since there are many feedback loops) and integrated (since the loops are all interwoven). Moreover, in stark contrast to the classic von Neumann computer architecture, most components are active most of the time.
6. Memory building and pattern recognition use the same mechanism, i.e., the feedback loop between the layers. On the one hand, "bottom up" memory building is self-acting, and to store some pattern it suffices to retain its order parameters. On the other hand, suppose there are stored parameters and some of the pattern's features are observed. The "top down" part of the feedback loop between the layers will then fill in the missing parts, until the dynamics have automatically restored the whole pattern (cf. Haken 2004, Section 17.1). In other words, synergetics offers an elegant, "combined" mechanism of memory building and information retrieval. Data compression and recovery are understood as a kind of feature extraction and pattern formation.[3]
7. More general, pattern formation is a particular kind of phase transition. Without a pattern, neural activity is incoherent, yet with a pattern it is orderly. Moving to the orderly state involves characteristic fluctuations, critical slowing down, and hysteresis that all have been observed in the motor arena (Haken 2006 (Chapter 11), Haken 2004 (Chapter 12), Haken 1996 (Part II), Haken, Kelso and Bunz 1985).
8. Different patterns are associated with distinct values of the order parameters. Here, too, typical oscillations can be observed, in particular if the sensory input supports several patterns .This effect can be demonstrated nicely with the help of flip-flop images, like Necker's cube (Haken 2004, Chapter 13). Difficult decisions seem to be similar: given a certain information basis, it may be hard to choose between several options (Haken 1996, Chapter 17), particularly if they are equally promising.

[3] The feedback mechanism nicely explains the well-known effect that recall is better, the more similar the situation when something was stored is to the situation when something has to be remembered: Each similarity serves as a clue that helps in restoring the whole pattern. Notice that this is quite different from "retrieving" a prepared bit of information from the memory of an ordinary computer.

Upstairs, downstairs

Let us summarize and extend the above argument within the more general framework. Our basic building blocks are modules, coping with a certain task. Moreover, they interact with their environment, i.e., they have an input and an output:

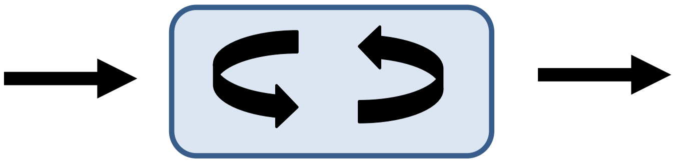

Illustration 16 (A single module, i.e., a functional unit)

Of course, given the overall hierarchical structure, a module may be further differentiated into interacting "sub-modules", etc. We will not go into these details, since the result of this train of thought is very similar to combining several modules into a "super-module" or a tier, consisting of (many) co-ordinated interacting modules, having an input as well as an output:

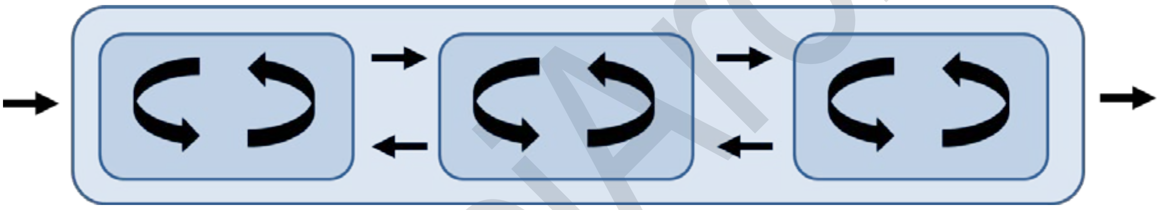

Illustration 17 (A single tier, and the modules it consists of)

Graphically, in the simplest case considered in the last section, there is a two-tier structure, with information flowing up and down, but also within the tiers:

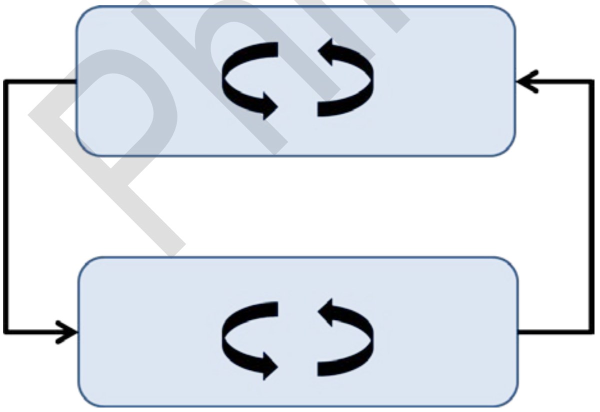

Illustration 18 (Two-tier structure with circular information flows)

If the locus of control is supposed to be rather "upstairs" than "downstairs", the output of the upper layer needs to be determined (at least mostly) by the state or, more generally, by the inner processes of that layer - and not by the input it receives from the lower tier. Thus the overall picture must be as follows:

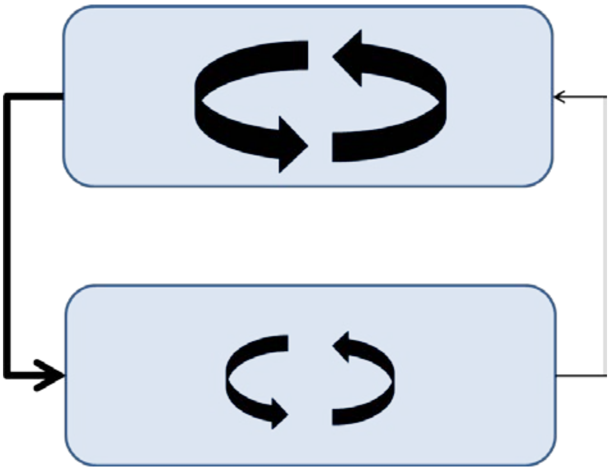

Illustration 19 (The same structure with more or less dominant feedback loops)

Self-organization and order parameters are a particularly elegant way to explain why top-down control is the rule and not the exception. In general, the overlying tier acts as a powerful 2nd order constraint that influences the subjacent layer much more than vice versa. If the bottom tier is the sensory realm, and the upper tier the cognitive realm, this corresponds nicely with the well-known view of Kant (1781) that "freedom, in the practical sense, is the independence of the will of coercion by sensuous impulses." If the uppermost tier is the self (see Illustration 15), "free will" is an appropriate subjective description for the (partial) "submergence" of the lower layers. To this end, language is an excellent tool since it provides explicit knowledge representation, concise chunks of information that may be combined in a transparent way, yielding resilient lines of argument that may lead to consistent action. It is no coincidence, that clear conceptual thinking and understanding, arguing, modelling and checking ideas are so important for us.

How strong is the loop within the upper layer? Roth (2003) remarks that a vast proportion of neurons in the associative cortex (up to 99%!) communicate with one another. Given this finding, some have concluded that we are constructivists, mainly revolving around ourselves, and building our own world. However, this conclusion is premature. First, in the course of evolution, those who forgot the outside world did not survive. The same result would occur if we could voluntarily influence the output of our sensory devices, i.e., perceive the world as we would like it to be. Second, the top level is thoroughly based on all the other layers below; it is not a "spirit in the sky". Third, the lower layers' input is still important if not decisive, if "informational updates" of the top layer are frequent (e.g., several times a second, say), and if this input influences the internal (circular) processes in the top level sufficiently. That is, in order to have a stable flow of information it would be straightforward if the circular processes in the top level reached an equilibrium without external input. However, if it is mainly the stimulus of the lower level(s) that gives them trend, such that the final result (i.e., the top down arrow on the left hand side in Illustration 19) may depend on the input in a crucial manner. (Although Lady Macbeth is whispering - not shouting - most of the time, she has a major influence on the overall plot.)

Constructivist ideas, e.g., the one just described, typically underestimate the influence of the "bottom up" input. However, if this flow of information breaks down completely, circular causality is destroyed, leaving the upper layer as a "free standing" (and thus failing) structure. Yet the above picture suggests a second kind of pathology: too much influence bottom up. In a well-organized mind there is healthy top down control, suppressing unwelcome impulses from lower tiers. Yet it is characteristic for many psychosomatic diseases that a lower tier has spontaneous impact on a higher tier. For example, panic attacks strike, that is, all of a sudden a person is flooded with fears. In the case of major depression, thoughts are negatively loaded and motivation fails. A major symptom of schizophrenia is uncontrollable sensory impressions, e.g., voices speaking up, non-existent persons coming into sight or more complex apparitions. If they issue commands, a sick person may do things he or she would never do in a healthy state of mind, and become a risk to himself and others.[4] It is also typical for these diseases, that they have episodes, i.e., there are sudden "bursts" that come and go. (The metaphor of the self being the rider of a horse is quite common. Here, it is a mulish horse that time and again threatens to unsaddle its rider.) For more psychopathological examples see Kelso and Tognoli (2009), p. 1112.

However, very often, activities are first located on the top tier and subsequently delegated further down. Learning some complex task, e.g., driving a car or playing a musical instrument, starts on the conscious level. One has to understand in great detail what kind of movement of limbs is required, when which movement is appropriate and how the arms and legs interact. Thus the proverb that all beginnings are difficult: they are slow and tedious and take enormous effort. Yet a major part of learning consists in *automatization*. Experienced drivers change gear without (conscious) thinking, and once a pianist has learned a musical piece their fingers know how to move. The saying that some faculty is "ingrained" or has become one's "second nature" catches perfectly what is going on: The skill is deeply rooted within the body, and control by the upper layer may be restricted to a bare minimum, e.g., a trigger. It is well known that typically (at least) ten years of thorough practice - 10,000 hours of training - are needed in order to learn some demanding activity "by heart". An engineer would say that this time is needed to replace (slow) software with (fast) hardware, i.e., neural networks have to be restructured and programmed in order to master some specialized task.[5]

[4] The cineaste autobiography of schizophrenic Nobel Laureate J. Nash, "A beautiful mind" (2001), succeeds in displaying this distorted worldview.

[5] Note that throughout this section, "locus of control" has been used in the sense of "who is being in charge", i.e., an entity, that rather issues commands than having to obey them. This is quite different to psychology's use of the term (Rotter 1966).

6) Nature’s bauplan

I am a strange loop (Hofstadter 2008)

In general, we have described and studied multi-layered (hierarchic), dynamical, self-referential and, to a large extent, also self-organizing information-processing systems, situated in a complex environment. A fully functional mind is a well-orchestrated, multi-modular organization; each and every part having its well-defined place and task, and being embedded in a multitude of loops. The overall result may be displayed in a single picture:

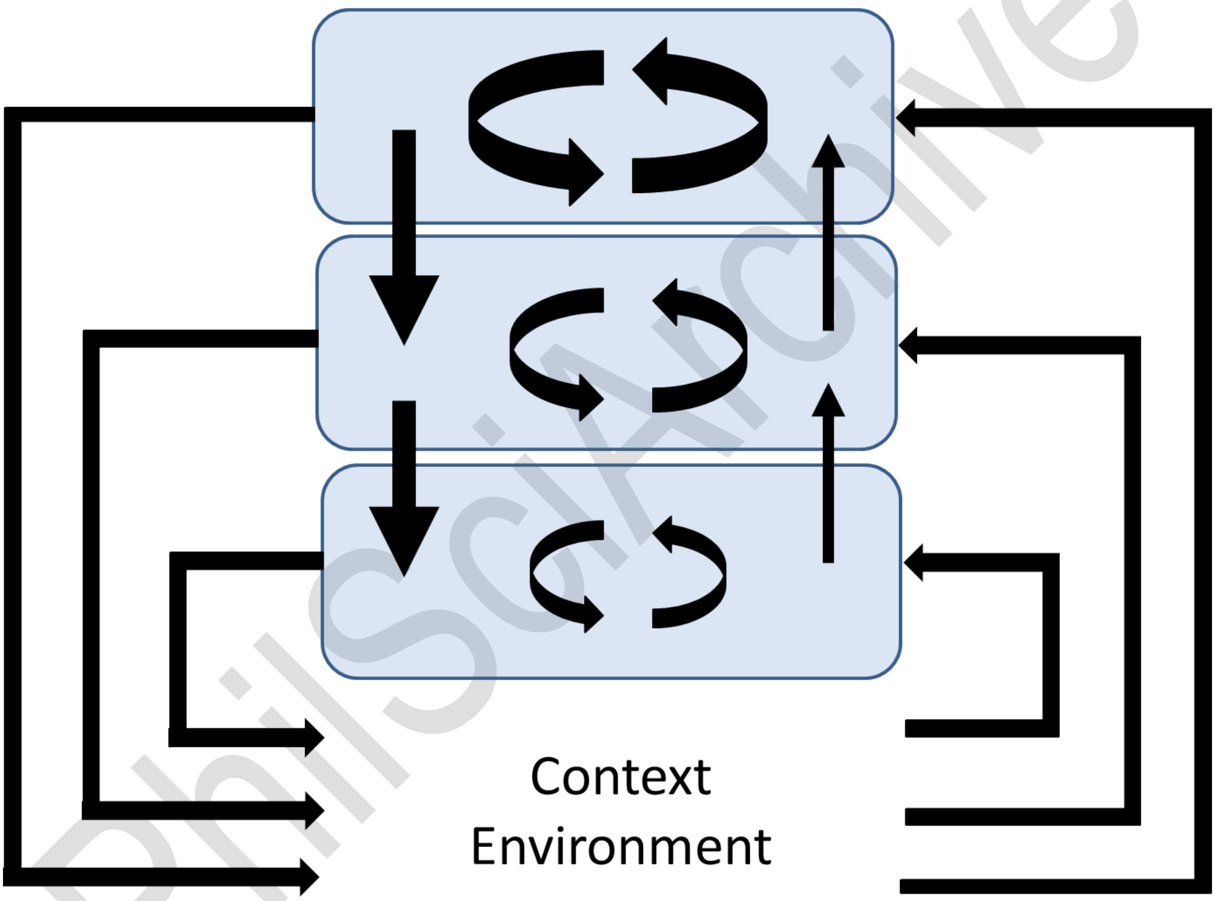

Illustration 20 (A complete multi-layered system with three kinds of feedback loops)

In a nutshell, the whole system consists of several layers. The locus of control is towards the top, that’s why the internal processes (⤿⮌) there are more important than those further down. The same holds for the interfaces between the layers: The influence “top down” is stronger than the influence “bottom up”, therefore the difference between 🡣 and 🡡. Moreover, there are sensorimotor loops. That is, all layers may cause actions (↦). If a certain action changes the environment of the system this change subsequently alters the sensory input (↤), possibly for all tiers.

Differentiation and integration

The dynamics of such a system can be characterized by **metastability**. Kelso and Tognoli (2009), pp. 105-108 explain: “One theory stresses that the brain consists of a vast collection of distinct regions ... The other school of thought looks upon the brain not as a collection of specialized centers, but as a highly integrated organ... metastability is an entirely new conception of brain organization ... Individualist tendencies for the diverse regions of the brain to express themselves coexist with coordinative tendencies to couple and cooperate as a whole. In the metastable brain, local and global processes coexist as a complementary pair, not as conflicting theories. Metastability, by reducing the strong hierarchical coupling between the parts of a complex system while allowing them to retain their individuality, leads to a looser, more secure, more flexible form of function ... No dictator tells the parts what to do. Too much autonomy of the component parts means no chance of coordinating them together. On the other hand, too much independence and the system gets stuck, global flexibility is lost.”

Supporting this point of view, Chennu et al. (2014) write: “Theoretical advances in the science of consciousness have proposed that it is concomitant with **balanced** cortical **integration and differentiation**, enabled by efficient networks of information transfer across multiple scales.” Thus numerical measures of dynamic complexity in general and for consciousness in particular have been proposed (Seth et al. 2011). Kelso and Tognoli (2009), p. 112, conclude “A delicate balance between integration (coordination between individual areas) and segregation (expression of individual behavior) is achieved in the metastable regime ... In a critical range between complete integration and complete segregation, the most favorable situation for cognition is deemed to occur ... measures of complexity reach a maximum when the balance between segregative and integrative forces is achieved.”

Using the idea of differentiation and simultaneous integration opens up another, rather straight road to self-awareness. In general, starting with a certain structure, new modules may evolve. Typically, they are first rather primitive but upon elaboration and segregation they obtain a certain “life of their own”. However, since the mental edifice is strongly interconnected, they are also readily integrated into the system already in existence. This happens spontaneously and on all levels:

Given a single sense organ, it is well-known that the first tier of sensory cells is occupied with restricted tasks (e.g., the direction of objects) and very limited areas (e.g., a certain spot of the visual field or a certain frequency of sound). Moving up the levels of analysis, the information is integrated, e.g., the areas of the visual field covered get larger and larger. Finally, all unimodal information is integrated into a comprehensive map, i.e., the world as we see, or hear, or smell, or feel it. The next “natural” step, of course, is to integrate all modalities into a comprehensive sensory model of the world, i.e., the world as we see, hear, smell *and* feel it.

Within a sufficiently rich perceptual model of the world, there is a prominent token: The self-image, i.e., a comprehensive map, representing the body in the perceptual realm. Combining this map with all the available information about the (inner) states of the body and motoric agency, readily yields a comprehensive body schema. When vocabulary and its accompanying structure evolve, a new module comes into being – the realm of language. Soon, within this area there is a pronounced token for oneself, the word "I", say. With the integration of language into the overall system, it is almost inevitable that the new word is connected with the existing body schema (the integrated representative of the entire body so far). This fusion creates a conceptually sharp and stabile distinction between oneself and the rest of the world. In other words, distinct self-awareness; an(other) emergent entity, further triggering the drastic effects described in the last paragraphs of Sections 2 and 4.

This train of thought underlines that it makes sense to distinguish between the protoself, the core self and the extended self, since each of them is located on a different mental tier. However, looking at the structure displayed in the rightmost column of Table 14, these selves also build on each other. More precisely, due to integration, the protoself is an integral part of the core self, the latter being a crucial part of the extended self. Finally, consolidating all available representations yields a single, comprehensive concept of oneself – one's self, embedded into a larger context (see Illustration 15). Quite similarly, Juarrero (2009) describes this process and its result as follows: "Dynamical closure always generates a boundary between the new emergent and the background. In the case of autopoietic structures the boundary is self-created by the very dynamics of the system. It can take the form of ... a dynamic phase separation between the emergent structure and the environment, or between the structure and its components."

Universal building blocks

Illustration 20 points out that, despite the nontrivial bauplan, there is a single universal building block, used over and over again: it is the information flow through *feedback loops*, also called "closed-loop causality" and "circular causality". This building block appears in various guises:

1. Sensorimotor loops, connecting the outside world with the internal mental life
2. Loops providing the information interchange between contiguous layers. In IT-jargon, the contiguous upper and lower layer acts as client and server
3. Loops within the tiers and modules, in particular loops serving as interfaces between modules, and loops mediating parts-whole relationships (e.g., between modules and their sub-structures)

Since evolution "likes" to reuse[6] approved building blocks, it is a straightforward conjecture that *all* information processing in natural, self-organizing information

[6] To "re-cycle" if you allow the play on words

systems is heavily based on feedback loops, from the processes within neurons, to small neural nets, neocortical columns, larger modules, cerebral areas, complete functional systems, the integral brain, and – finally -- the whole nervous system.

Moreover, every biological unit, but also every robot, is situated in some context. Thus the very first loop, that is, the elementary sensorimotor loop connecting the "machine" with the outside world, is inevitable. It fits well into our understanding of evolution that this very first loop was re-used, modified (split, differentiated, put to a different use, redirected, strengthened, weakened, dissolved etc.), and gradually extended. Thus creating specialized modules, distinct areas and hierarchic layers, all of them tightly linked, and combined in an overall sound architecture, "thinking" (more and more sophisticated internal information processing) developed. Finally, well-orchestrated mental edifices with a clear understanding of themselves and their situatedness appeared.

The existing literature puts much emphasis on "downward determination" and "circular causality". According to the reasoning in this contribution, these terms are important, however, they may also easily miss their target. First, the best formal account of causality is based on directed, *acyclical* graphs (Pearl 2009). Second, although it is correct to acknowledge the role of top-down processes (giving the higher layers at least some influence), one should not overlook the fact that each of them is just a leg of more important information-processing loops. The same with the idea of a "closed loop". Of course, by definition, every loop is closed, i.e., the end-point of some process coincides with its starting point. However, there is also contextual input and procedural output which may be crucial. In this sense, information processing loops are open, they interact massively with one another and their environment. Third, causation and determination are often contrasted with chance and freedom. Since there is an abundance of reasons and causes and since, traditionally, free will has been associated with non-determinism, one is easily led down the primrose path of fundamental discussion.

This author thinks that dynamical system's theory should take centre stage, as its emphasis is on the behaviour of complex systems. Thus it is preoccupied with systems being composed of many particles and being held together by "forces" of all kinds. Moreover, context and constraints play a major role, and one has to consider numerous and diverse factors, be they deterministic or stochastic. The modes of such systems range from straightforward convergence, and (quasi-)deterministic behavior to arbitrary random fluctuations with all kinds of regularity and irregularity in-between (e.g., periodicity, more or less stable attractors, turbulence and chaos). There also seems to be self-organized criticality (Èrdi 2008, Sornette 2009), in particular when a certain state of mind (in this view a certain attractor) becomes unstable due to saturation, self-amplification (Haykin 2013, p. 442-443) and resonance, e.g., when the best fitting option supersedes all others. Several authors remark that the brain seems to be working "close to instability points" or "at the edge of chaos" (e.g., Legenstein and Maass 2007), when information throughput and complexity are highest. A thought-provoking application of these ideas to our subject can be found in Andrade (2008) who sees a hierarchy of regimes: Physical Information Systems, Information Gathering and Using systems, and Hierarchical Dynamical Information Systems.

What is crucial is the flow of information. This flow is organized in myriads of feedback loops, all of them working simultaneously but at the same time being heavily (hierarchically) interconnected. In the style of Swift's society of fleas, a loop has smaller loops on which it relies and still larger loops that build on it. In addition, this massively parallel, "Goldilocks-like"[7] and "metastable" (Kelso and Tognoli 2009) processing of information is dynamic: It always changes, never converges or comes to an end. Instead, at any one time, there is some amount of activity which is also variable. However, although the content and the intensity of the internal course of events alter ceaselessly, and, at times, almost unpredictably (e.g., due to new input), the mental stream is kept within certain bounds. In a deep sense, thinking is like (endless) weaving, with elementary mesh loops combining into patches, models and cloth. That's the bottom-up view. However, at the same time there is "downward causation." That is, the whole "loom" (i.e., the entirety of all meshes) and the patterns it produces blaze the trail for subsequent activities on lower tiers.

7) Construction principles

It is perhaps worth pointing out that our analysis predicts the possibility of constructing a conscious artifact and outlines some key principles that should constrain its construction (Tononi and Edelman 1998)

According to Knuth (2011), programming is an art. Thus it can be concluded from the above that natural evolution is a great artist. What about us? How far have we come, what's the state of (our) art? We know how to build computers. More generally, we have learned to build fast, reliable and competitive hardware. For example, although the capacity of the human memory is incredibly large, cutting-edge technical storage devices reached the same dimension. In other areas, in particular speed, modern semi-conductor micro-technology has already left its biological counterparts far behind.

With respect to software we know how to "knit" programs, we are able to create flexible modules, we can manage multi-tier architectures of considerable size, and there even is a stable network connecting millions of devices worldwide. In other words, software-engineering is much less of an art than it was decades ago. Nowadays, it is routine work to design, to implement and to run large software packages on distributed IT-systems.

[7] not too tight, not too loose, cf. Juarrero (2009)

However, looking at robots, progress has been slow. Traditional artificial intelligence and cognitivist approaches put much emphasis on symbolic manipulation, modelling, and planning, which all come with a heavy computational burden.[8] They produced large programs that did not really take embodiment and situatedness into account, and all failed the test of practice, i.e., everyday situations that are both fuzzy and complex.

Learning from nature

Coming from neuropsychology, Hebb (1949) found the simple and general learning rule that “what fires together wires together”, i.e., that the association between neurons is strengthened when they are commonly active. He thus founded the research area of neural networks which, ever since, has fluctuated between hype and neglect. Extending Hebb’s idea, Rosenblatt (1958) introduced associative memory, i.e., the idea that information is stored in the wiring between the neurons. More precisely, given a set of neurons, it is the strength of the links between the neurons that constitutes the material basis of memory. Depending on the input to such an array of interconnected neurons (a module), a large number of outputs, i.e., patterns are possible. First, neuronal nets were trained “directly”. Later, supervised learning was extended and superseded by backpropagation, reinforcement learning and completely unsupervised learning, leading to self-organizing nets (Kohonen 2001).

However, despite the biological start and motivation, and although almost every publication refers to information processing in the brain, research has increasingly focused on formal aspects. Today, neural networks are mainly used as a particular tool in mathematical optimization and statistical learning (Du and Swamy 2014). This has slowed down progress, which always happened when the field neglected its empirical, in particular biological, roots. Thus, although it is somewhat astonishing and shows the immense potential of neural nets that they do well in arenas far away from the challenges where they originally succeeded, it is to be expected that mainstream research fared much better if it focused on the tasks of real-life (cf. some chapters, in particular the last one, in Haykin 2013). These are many-faceted, often vague and error-prone, very dynamical, involve sensory and motoric tasks, and have to be solved in real time. Ciresan et al. (2012) give an example: “The human visual system efficiently recognizes and localizes objects within cluttered scenes... Deep hierarchical neural models roughly mimic the nature of mammalian visual cortex, and are among the most promising architectures for such tasks.”

The ultimate challenge consists in building an autonomous machine endowed with a self-extracting multi-layered control system, i.e., to create a mentally developing robot (e.g., Weng 2004, 2007; Cangelosi and Schlesinger 2015). To this end, it seems helpful to ask how nature succeeded in programming its “survival machines”

[8] For an instructive overview and a comparison see Vernon (2010) and Shapiro (2010).

(Dawkins 2006). We have already mentioned her massive parallel approach. Ceaseless as these innumerate processes may be, computation costs time and effort (energy, resources, etc.). Therefore a **first maxim** must be to minimize this expense.

Haken (2004), p. 17, gives a nice example: "It is often believed that in [the] recognition process an enormous number of details are analysed … The evolutionary process suggests the opposite." More generally, it seems appropriate only to "think" as much as necessary in order to get a desired result. In other words, elegant solutions restrict central information processing to the inevitable minimum and take advantage of the physics of the body as well as the services of the environment whenever possible. More precisely, nature's economical recipe seems to "shift the computational load from the [central] controller to the morphology and physical properties of the embodiment … The controller is challenged to maximally exploit the physical peculiarities of the body in its interaction with the environment." (Der and Martius 2011, pp. 29).

Efficient management of a robot uses the latter's scarce resources optimally, that is, it externalizes burdens whenever possible, maximizing the attainable effect but minimizing internal costs. Paradigm examples can be found in Der and Martius (2011). Crucial ideas are collected in Brooks (1999) who underlines that it is embodiment that provides meaning (semantics), that a successful robot needs extensive front and back ends (i.e., powerful sensory and motoric devices), that the world is its own best model, and that intelligence is rather determined by the dynamics of the interaction with the world than by explicit representation and reasoning.

The **second maxim** is to start with simple building blocks and to use them time and again, tailoring them to some specific need. Ubiquitous feedback loops connect the individual with the outside world, but also assemble neurons into small, big and huge units – from neocortical columns to brain hemispheres. Although these units' structures cannot be identical – since they have to cope with different problems - they all work according to similar principles, and need to be integrated if necessary. For example, it is well known that pattern formation is almost identical to pattern recognition, and similar to decision making (see numerous references to Haken throughout this contribution). Visual and auditory perception are "a tale of two sides" (Haykin and Chen 2007). Moreover, temporal binding of brain areas always depends on spontaneous synchronization.

In a nutshell, there are myriads of modules, all acting in parallel and simultaneously. The formidable task of fine-tuning is mostly solved via hierarchic and dynamic self-regulation, channeling the flow of information. Since timing is crucial, so are "spike trains" (Gerstner and Kistler 2002) and their precise synchronization (Haken 2008). Memory is also organized in a unified way, with content being distributed throughout a net of neurons rather than put in a single "drawer" at a particular location. With the information being laid down in the matrix of neuronal connections, memory is dynamic and self-organizing, with some input evoking a certain dynamic response, typically resulting in a fitting output.

In this view, the basic functional unit is a module, i.e., an array of connected neurons. It is rather obvious that it may be programmed in two completely different ways: On the one hand, there is “normal” plasticity. Upon gradually strengthening or weakening the connections within this group of neurons, memory or any other function changes slowly. However, on the other hand, there is also “fast learning”, especially during sensitive phases. A plausible mechanism to this end is “massive pruning”, i.e., to start with a large number of neurons and links, and subsequently eliminating most of them during the learning process, resulting in hardware that has been customized quickly to a certain context.

These completely different ways of putting a module into operation may explain the enormous differences upon learning a similar task, e.g., between first and second language acquisition. Due to the first maxim, i.e., since it is costly to first build a large field of neurons and then destroy most of it, the later process should be the exception in normal (adult) life. However, when the focus is on rapid development, i.e., in children, the second process should be widespread, and explains in part why they need such an enormous amount of energy to build up their mental edifice.

Since learning is tedious (consuming time and energy), one can also expect that nature uses prior information whenever available. That is, pre-structured neuronal networks, ready-made for a specific task, should be ubiquitous. On such a basis, learning rather resembles grouting and fine-tuning than a major effort which is necessary when building a structure from scratch. The example of language acquisition demonstrates the enormous difference: Within a short sensitive phase, children learn to master their mother tongue better than adolescents do a second language. Moreover, hardly any amount of training after the sensitive phase will suffice to reach the level of command a child has obtained in passing.

The **third maxim** is to use self-organization wherever possible. That is, instead of teaching a robot many special tricks or having him store one sensory impression after the other explicitly, it seems much more advisable to compress the necessary information to a bare minimum and re-establish the original when necessary. For example, the popular format MP3 does not store a song completely. Rather, it stores and compresses the information relevant to the human ear, ignoring the rest. It is also not necessary to put down an image completely. Rather, it suffices to retain some particular features and fill in the rest upon request, i.e., given certain clues. The human eye is also not a camera taking picture after picture and combining them into a movie. Rather, elementary eye saccades look for differences and just update those parts of our view that have changed.

Haken (2006), p. 28, summarises: “Quite often it is assumed that the incoming pattern is compared with templates. However, the storage of a template would require quite a large amount of information. Therefore, one might imagine, in the sense of synergetics, that only specific characteristic features are stored in the form of order parameters which may then be called upon to generate a detailed picture. In this sense then, pattern recognition becomes an active process in which new patterns are formed in a self-organized fashion…"

It may be added that every conventional computer program can be understood as a compact recipe to some end. Upon its execution, that is, upon putting it in a certain environment, it is decompressed and creates all the effects it is supposed to produce. Interestingly enough, Turing showed that very few building blocks (in particular loops and bifurcations) suffice to compute anything that is calculable. Notice the deep-rooted similarity: Computer programs, genes, inseminated egg cells - indeed any kind of offspring - are seeds that, if put into an appropriate context, develop rather automatically, they "unfold" there so to speak. However, self-organization goes much further.

First, due to the permanent feedback of the organism and its environment, self-development is strongly adaptive in the sense that the course of "unfolding" is very much guided by local, specific boundary constraints. For example, given the initial competence of language acquisition, every healthy child is able to learn any language perfectly, just depending on the area where it grows up. In the extreme, the context acts like cladding being filled with the evolving structure.

Second, development is automatic and follows general rules: It always starts with crude, rather rudimentary beginnings. Given a reasonable context, humble abilities differentiate into sophisticated ones. An appropriate amount of guidance and protection certainly helps, however, most of the construction work has to be delivered by the developing structure. Moreover, depending on the ability to be acquired, there are more or less restricted time slots. Typically, it is much easier to learn a skill earlier in life, when the brain and the body are "made for" the acquisition of new faculties of all kind. Since abilities typically build on each other, there is also a natural order in which skills should be acquired. It is futile to teach mathematical subtleties when the pupils have not yet understood elementary numbers.

Third, despite all the work that is going on, upon gradually extending the system "loop by loop", the whole system remains robust. One could call this "self-organized stability." New modules are established, tested, run in, geared to each other, gradually added to the whole system, and finally used on a regular basis. Again, in a quite self-organized manner, single building blocks form larger structures, until, when the system has matured, all layers are "installed and ready." Trying to design and implement a complete software edifice for a robot thus seems a hopeless endeavor. Instead, nature chose not to build "Rome" in a day, but have humble beginnings grow and thrive.

Fourth, with complex dynamic systems come all kinds of emergent phenomena. In particular, larger aggregates attain abilities that their components do not have. For example, single neurons have a very limited behavioral repertoire, yet neuronal nets can store information and compute complex functions. The components of a cell are just biochemistry, yet the cell can replicate, i.e., manufacture a copy. Multicellular organisms can differentiate, forming versatile bodies with astonishing features." Such "phase transitions" when "completely new dimensions" are reached, are not the exception, but the rule. They happen quite often, appear spontaneously and all of a sudden. The popular idea of *self-organized criticality* (SOC) even suggests that evolving systems may have – or attain - the ability to provoke such "tipping points" (see the vast literature inspired by Bak et al. (1987)). Typically, the new properties are almost unpredictable and, at best, explainable with the wisdom of hindsight. Nevertheless, they may have dramatic consequences. The amazing phenomenon of self-awareness fits perfectly well into this global picture.

8) Synthesis

Life isn't about finding yourself. Life is about creating yourself (attributed to G.B. Shaw).

In a nutshell, conceptually clear self-awareness, combined with far reaching autonomy, seems to emerge quite straightforwardly in self-referential, multi-layered dynamical information processing systems endowed with a rich perceptual image and a complex language. The details may be tricky and not yet fully understood.

However, many crucial concepts and mechanisms already have been detected. The computer sciences and neurosciences, but also physics, biology, and psychology as well as the philosophy of mind all contribute to the understanding of our mental equipment. Although they have different perspectives and use distinct methods, they do not contradict but rather help each other to create a comprehensive picture. This article tried to demonstrate that these lines of investigation are converging towards a complete theory of the human self. Although it has been a long journey, the route travelled can be summarized in a few steps:

1. **Embodiment**. Fundamentally, it is "no body, never mind" (Damasio 1995). Only with embodiment come situatedness, real-life problems, and the circular flow of information (Thelen and Smith 1994, Brooks 1999, Gallagher 2006, Pfeifer and Bongard 2007)
2. **Feedback loops** are the ubiquitous, versatile and dynamic functional building blocks of all mental life, reused and reshaped by evolution over and over again. Development is due to the body's permanent active interaction with its environment. (Section 6)
3. **Progressive differentiation and integration.** Piaget's universal mechanism explains how a modest starting point can develop into a complex mental construction. That is, feedback loops used as "lassos" are able to capture

features, insights, and abilities - one by one - resulting in reliable growth, i.e., gradual differentiation within an integrated total system, thus solving the stability-plasticity dilemma (Richardson and Thomas 2008, Smith 2009).

4. **Self-organizing neural networks** are the physical building blocks used by nature. Since, from a computational point of view, recurrent neural networks are immensely powerful (Haykin, Section 15.5.) they are able to support a tremendous mental edifice. Moreover, their repetitive "goldilocks-like" architecture is adaptive, and spontaneously develops into massively parallel, hierarchical structures (nets, columns, modules, layers etc.).
5. **Evolution toward complexity**. Out of basic sensorimotor loops thus evolved more advanced functions, in particular associative memory, learning and meta-learning (Haykin 2013, in particular p. 864), abstract knowledge and reasoning (Cangelosi and Schlesinger 2015, Chapter 8), and language (Asada 2015). Schlesinger (2009), p. 192, summarizes: "There is a broad community of researchers who agree that sensorimotor activity is a fundamental starting point for cognition, in general." Along this track, primitive neuronal networks also became more and more complex, finally forming something like a triune brain (MacLean 1990), consisting of three major tiers: the reptilian complex (being mainly located in the brain stem, controlling homeostasis and generating instincts), the paleo-mammalian complex (being located in the limbic system, bringing about emotions), and the neomammalian complex (located in the neocortex, and hosting cognition in a broad sense).
6. **Dynamic system theory.** The human brain is the most complex structure known. Therefore the terms and ideas of dynamical system theory are useful. For example, spontaneous pattern formation is associated with appropriate order parameters, an idea stressed by synergetics. The re-utilization of the same building blocks (neural networks, feedback loops, differentiation and integration) on different scales results in an overall fractal structure.[9] More generally, since the architecture of many systems is closely related to their purpose (Rubenstein and Rakic 2013), in evolution as well as systems' design, form follows (from) function, or even "function becomes structure." Altogether, it is straightforward to conceive of the brain as a self-organized synergetic computer (Kelso 1995, Haken 2004).
7. **Development** also follows the general logic of dynamical system theory, i.e., it is both spontaneous and self-organized (no contradiction). All that is needed is a germ that will "self-inflate" if put in a suitable environment. That is, via permanent interaction such a system will differentiate and move from primitive to complex, thereby reaching a number of "bifurcation points" and passing through a series of critical phases. Quite characteristically, during this process, rather stable stages alternate with rather chaotic transitions, and with every major evolutionary step, the system all of a sudden acquires/demonstrates new, emergent properties, some of them being truly surprising and almost unpredictable (see e.g., Thelen and Smith 1994, Ritter et al. 2011, Cangelosi and Schlesinger 2015, in particular p. 7, and many contributions of Haken).

[9] For a striking example see McClelland and Vallabha (2009), pp. 16.

8. **Representations of oneself.** Due to the multi-layered structure, there are several sensory-based descriptions of the environment available. These are integrated into a perceptual “world model”. Owing to circularity, there are also tokens for oneself (one’s body, or at least parts of it) in each of the sensory realms which are integrated into one body image. When combined with motoric agency, they form a complete body schema.
9. With the realm of **language** evolves a second, conceptually precise description of the world. Moreover, within this realm, there emerges a concept of oneself. There is now a word for myself. Combining this word and the body schema triggers an “autocatalytic reaction” (described in detail in Section 1), leading to a complete reorganization of the mental realm with an integrated, conceptually clear token of / for oneself on top. Thus man has become “ego-conscious”; he has and is a personal self (Saint-Mont 2001). However, due to the multi-layered architecture, several forms of consciousness can be distinguished (Donald 2002, Damasio 2010).
10. **Free will.** It is well-known that layers building on each other operate on different time scales (e.g., Juarrero (2009), p. 99; Newell et al. 2009, and the references given there). As a rule, the lower the layer, the faster it works (just compare representative physical, chemical, biological and social processes). However, it is also well known that, typically, the upper layers are in control, i.e., the system’s hierarchy also serves as a command structure. Thus the uppermost layer becomes the most natural locus of control for the whole system which corresponds nicely with the classical view of a rather autonomous personal self, being the charioteer of mind and body.[10]

[10] It may be added that given this standpoint, Libet’s results can be interpreted in an elegant way: The basic sensorimotor tier quickly sets a behavioural default. Bottom-up, this fixing appears in the conscious mind as a decision, although, in this case, it is just an a posteriori rationalization. However, operating on a slower time scale, but being truly in charge, top-level consciousness may readily overrule the lower tier’s move (e.g., Donald 2002, Bandura 2008, Baumeister 2010).

References

Abbott, E.A. (1884/1992). Flatland. A romance of many dimensions. New York: Dover Publ.

Anderson, M.L. (2003). Embodied cognition: a field guide. Artificial Intelligence 149, 91-130.

Anderson, P.W. (1972). More is different. Broken symmetry and the nature of the hierarchical structure of science. Science 177, 393-396.

Andrade, E. (2008). Remarks on Collier. Cybernetics and Human Knowing 15(3-4), 87-99.

Asada, M. (2015). Toward language: Vocalization by cognitive developmental robotics. Chapter 10 in Cheng (2015), 251-273.

Asada, M.; MacDormann, K.F.; Ishiguro, H.; and Y. Kuniyoshi (2001). Cognitive developmental robotics as a new paradigm for the design of humanoid robotics. Robot. Auton. Syst. **37(2-3)**, 185-193.

Asada, M.; Hosoda, K.; Kuniyoshi, Y.; Ishiguro, H.; Inui, T.; Yoshikawa, Y.; Ogino, M.; and C. Yoshida (2009). Cognitive developmental robotics: a survey. IEEE Trans. Auton. Mental Develop. 1**(1)**, 12-34.

Baer, J.; Kaufman, J.C.; and R.F. Baumeister (eds., 2008). Are we free? Psychology and free will. Oxford University Press.

Bak, P.; Tang, C; and K. Wiesenfeld (1987). Self-organized criticality: an explanation of 1/f noise. Physical Review Letters 59 (4), 381–384.

Bandura, A. (2008). Reconstrual of "free will" from the agentic perspective of social cognitive theory. Chapter 6 in: Baer et al. (2008), 86-127.

Baumeister, R. (2010). Understanding free will and consciousness on the basis of current research findings in psychology. Chapter 3 in: Baumeister, R.; Mele, A.; and K. Vohs (2010, eds.). Free will and consciousness: How might they work? Oxford University Press.

Barsalou, L.W. (2008). Grounded cognition. Annu. Rev. Psychol. 59, 617-645.

Bermúdez, J.L. (2014). Cognitive science (2nd ed.). Cambridge Univ. Press.

Brooks, R.A. (1999). Cambrian intelligence. The early history of the new AI. Cambridge: MIT Press.

Bunge, M. (2003). Emergence and convergence: qualitative novelty and the unity of knowledge. Toronto: University of Toronto Press.

Campbell, D.T. (1974). `Downward causation' in hierarchically organised biological systems. In: Ayala, F.J., and T. Dobzhansky (eds.). Studies in the philosophy of biology. Berkeley and Los Angeles: University of California Press, 179-186.

Cangelosi, A.; and M. Schlesinger (2015). Developmental robotics. From babies to robots. London: MIT Press.

Cheng, G. (2015, ed.). Humanoid robotics and neuroscience. Science, engineering and society.Boca Raton: CRC Press, Taylor & Francis.

Chennu, S.; Finoia, P.; Kamau, E.; Allanson, J.; Williams, G.B.; Monti, M.M.; Noreika, V.; Arnatkeviciute, A.; Canales-Johnson, A.; Olivares, F.; Cabezas-Soto, D.; Menon, D.K.; Pickard, J.D.; Owen, A.M.; and T.A. Bekinschtein (2014). Spectral Signatures of Reorganised Brain Networks in Disorders of Consciousness. PLoS Comput Biol 10(10): e1003887. doi: 10.1371/journal.pcbi.1003887.

Ciresan, D.C.; Meier, U., Masci, J.; and Jürgen Schmidhuber (2012). Multi-column deep neural network for traffic sign classification. Neural Networks 32, 333-338.

Damasio, A. (1995). Descartes' error: emotion, reason and the human brain. Picador.

Damasio, A. (1999). The feeling of what happens. Orlando: Harcourt.

Damasio, A. (2010). Self comes to mind. New York: Pantheon Books.

Dawkins, R. (2006). The selfish gene. 30 Anniversary edition. Oxford: Oxford University Press.

De Preester, H.; and V. Knockaert (2005, eds.). Body image and body schema. Interdisciplinary perspectives on the body. Amsterdam: John Benjamin.

Deacon, T.W. (1997). The symbolic species: The co-evolution of language and the human brain. London: Penguin.

Deacon, T.W. (2007). Three levels of emergent phenomena. In: Murphy, N.C.; and W.R. Stoeger (eds.). Evolution & emergence: Systems, organisms, persons. Oxford: Oxford University Press, 88-110.

Der, R.; and G. Martius (2012). The playful machine. Springer.

Donald, M. (2002). A mind so rare. New York: Norton.

Du, K.-L.; and M.N.S. Swamy (2014). Neural networks and statistical learning. London: Springer.

Èrdi, P. (2008). Complexity explained. Springer.

Feynman, R. (1988). What I cannot create, I do not understand. Note on a Caltech blackboard. www.onionesquereality.files.wordpress.com/ 2008/12/ feynman_blackboard5.jpg

Fodor, J. (1983). The modularity of mind. Cambridge, MA: Cambridge Univ. Press.

Gallagher, S. (2006). How the body shapes the mind. Oxford University Press.

Gallup, G. G., Jr. (1970). Chimpanzees: self recognition. Science 167, 86-87.

Gerstner, W.; and W.M. Kistler (2002). Spiking neuron models. Single neurons, populations, plasticity. Cambridge: Cambridge University Press.

Haken, H. (1996). Principles of brain functioning. A synergetic approach to brain activity, behaviour and cognition. Berlin: Springer.

Haken, H. (2002). Brain dynamics. Synchronization and activity patterns in pulse-coupled neural nets with delays and noise. Berlin: Springer.

Haken, H. (2004). Synergetic computers and cognition. A top-down approach to neural nets (2nd ed.). Berlin, Heidelberg: Springer.

Haken, H. (2006). Information and self-organization. A macroscopic approach to complex systems (3rd ed.) Berlin: Springer.

Haken, H. (2008). Brain dynamics. An introduction to models and simulations (2nd ed.) Berlin: Springer.

Haken, H.; Kelso, J.A.S.; and H. Bunz (1985). A theoretical model of phase transitions in human hand movements. Biol. Cybernetics **51**, 347-351.

Haken, H.; and G. Schiepek (2010). Synergetik in der Psychologie (2nd ed.) Hogrefe-Verlag.

Hauser, M.D.; Chomsky, N.; and W.T. Fitch (2002). The faculty of language: what is it, who has it, and how did it evolve? Science 298, 1569-1579.

Haykin, S. (2013). Neural networks and learning machines (3rd ed.). Pearson.

Haykin, S.; and Z. Chen (2007). The machine cocktail party problem. Chapter 3 in Haykin et al. (2007), 51-75.

Haykin, S.; Príncipe, J. C.; Sejnowski, T. J.; and J. McWhirter (eds., 2007). New directions in statistical signal processing: From systems to brain. Cambridge (MA) & London: MIT Press.

Hebb, D.O. (1949). The organization of behavior. A neuropsychological theory. New York.

Hofstadter, D.R. (2008). I am a strange loop. Basic books.

Juarrero, A. (2009). Top-down causation and autonomy in complex systems. In: Murphy, N et al. (eds.), 83-102.

Kane, R. (ed., 2011). The Oxford handbook of free will (2nd ed.) Oxford University Press.

Kant, I. (1781). The critique of pure reason. Translated by J. M. D. Meiklejohn. http://www.gutenberg.org/ebooks/4280.

Kelso, J.A.S. (1995). Dynamic patterns: The self-organization of brain and behavior. Cambridge (MA): MIT press.

Kelso, J.A.S.; and E. Tognoli (2009). Toward a complementary neuroscience: Metastable coordination dynamics of the brain. In: Murphy, N et al. (eds.), 103-124.

Knuth, D.A. (2011). The art of computer programming (3rd ed.) Addison Wesley.

Kohonen, T. (2001). Self organizing maps (3rd ed.). Berlin, Heidelberg: Springer.

Kornhuber, H.H.; and L. Deecke (1965). Hirnpotentialänderungen bei Willkürbewegungen und passiven Bewegungen des Menschen: Bereitschaftspotential und reafferente Potentiale. Pflügers Archiv Physiologie 284, 1-17.

Kuzawa, C.W.; Chugani, H.T.; Grossman, L.I.; Lipovich, L.; Muzik, O.; Hof, P.R.; Wildman, D.E.; Sherwood, C.C.; Leonard, W.R.; and N. Lange (2014). Metabolic costs and evolutionary implications of human brain development PNAS 111 (36), 13010-13015.

Legenstein, R.; and W. Maass (2007). What makes a dynamical system computationally powerful? Chapter 6 in Haykin et al. (2007), 127-154.

Lewin, K. (1939). Field theory and experiment in social psychology. American Journal of Sociology 44 (6): 868–896.

Libet, B. (1985). Unconscious cerebral initiative and the role of conscious will in voluntary action. Behavioral and Brain Sciences 8, 529-566.

MacLean, P.D. (1990). The triune brain in evolution: role in paleocerebral functions. New York: Plenum Press.

Mareschal, D.; Leech, R.; and R.P. Cooper (2009). Combining connectionist and dynamic systems principles in models of development: The case of analogical completion. Chapter 10 in Spencer et al. (2009), 203-217.

Marr, D. (1971). Simple memory: A theory of archicortex. Phil. Trans. of the Royal Soc. of London, Ser. B, **262**, 23-81.

Marr, D. (1982). Vision: A computational investigation into the human representation and processing of visual information. San Francisco: W.H. Freeman.

McClelland, J.L.; and G. Vallabha (2009). Connectionist models of development: Mechanistic dynamical models with emergent dynamical properties. Chapter 1 in Spencer et al. (2009), 3-24.

Murphy, N.; Ellis, G.F.R., and T. O'Connor (2009). Downward causation and the neurobiology of free will. Berlin, Heidelberg: Springer: Complexity.

Newell, K.M.; Liu, Y.-T.; and G. Meyer-Kress (2009). Timescales of changes in connectionist and dynamical systems approaches to learning and development. Chapter 6 in Spencer et al. (2009), 119-138

Pearl, J. (2009). Causality (2nd ed.) Cambridge: Cambridge University Press.

Pfeifer, R.; and C. Scheier (1999). Understanding intelligence. Cambridge, MA: MIT Press.

Pfeifer, R.; and J. Bongard (2007). How the body shapes the way we think: A new view of intelligence. Cambridge, MA: The MIT Press.

Richardson, F.M.; and M.S.C. Thomas (2008). Critical periods and catastrophic interference effects in the development of self-organizing feature maps. Developmental Science, 11(3), 371-389.

Ritter, H.; Haschke, R.; Röthling, F.; and J. Steil (2011). Manual intelligence as a Rosetta Stone for robot cognition. Robot. Res. 66, 135-146.

Rochat, P. (2003). Five levels of self-awareness as they unfold early in life. Consciousness and Cognition 12, 717–731.

Rosenblatt, F. (1958): The perceptron: a probabilistic model for information storage and organization in the brain. Psychological Reviews 65, 386-408.

Roth, H. (2003). Fühlen, Denken, Handeln: Wie das Gehirn unser Verhalten steuert. Frankfurt: Suhrkamp.

Rotter, J.B. (1966). Generalized expectancies for internal versus external control of reinforcement: Psychological Monographs: General & Applied 80(1), 1-28.

Rubenstein, J.L.R.; and P.R. Rakic (2013). Comprehensive developmental neuroscience: Neural circuit develolpment and function in the healthy and the diseased brain. Amsterdam: Academic Press.

Saint-Mont, U. (2001). Das Gehirn und sein Ich. Über die Evolution und Konstruktion des Bewußtseins. Berlin: Verlag für Wissenschaft und Bildung.

Schaller, S. (1991). A man without words. New York: Summit Books.

Schlesinger, M. (2009). The robot as a new frontier for connectionism and dynamic system theory. Chapter 9 in Spencer et al. (2009), 182-199.

Searle, John (1980). Minds, brains and programs. Behavioral and Brain Sci. 3 (3), 417–457.

Sellars, R.W. (1970). Principles of emergent realism: The philosophical essays of Roy Wood Sellars. In: W. Preston Warren (ed.) St. Louis: Warren H. Green.

Seth, A.K.; Barrett, A.B.; and L. Barnett (2011). Causal density and integrated information as measures of conscious level. Phil. Trans. R. Soc. A **369**, 3748-3767.

Shapiro, L. (2010). Embodied cognition. London and New York: Routledge.

Singer, W. (2007). Binding by synchrony. Scholarpedia, 2(12):1657.

Smith, L.B. (2009). Dynamic systems, sensorimotor processes, and the origins of stability and flexibility. Chapter 4 in Spencer et al. (2009), 67-85.

Sornette, D. (2009). Critical phenomena in natural sciences: Chaos, fractals, selforganization and disorder: Concepts and tools. 2nd ed. Springer.

Souza Vieira, F. de; and C.H. El-Hani (2008). Emergence and downward determination in the natural sciences. Cybernetics and Human Knowing 15(3-4), 101-134.

Spencer, J.P.; Thomas, M.S.C.; and J.L. McClelland (2009, eds.). Toward a unified theory of development. Connectionism and dynamic system theory re-considered. New York: Oxford University Press.

Sperry, R.W. (1983). Science and moral priority: merging mind, brain, and human values. New York: Columbia University Press.

Thelen, E.; and L.B. Smith (1994). A dynamic systems approach to the development of cognition and action. Cambridge MA: MIT Press – Bradford Books.

Tononi, G.; and G.M. Edelman (1998). Consciousness and complexity. Science 282, 1846-1851.

Van Gulick, R. (1995). Who's in charge here? And who's doing all the work? In: Heil, J.; and A. Mele (eds.). Mental causation. Oxford: Clarendon Press, 233-256.

Vygotsky, L. (1978). Mind in society: The development of higher mental processes. Cambridge, MA: Harvard University Press.

Watson, J.B. (1913). Psychology as the behaviorist views it. Psych. Review 20, 158-177.

Wegner, D. M. (2002). The illusion of free will. Cambridge: MIT Press.

Weng, J. (2004). Developmental robotics: theory and experiments. International Journal of Humanoid Robotics 1(2).

Weng, J. (2007). On developmental mental architectures. Neurocomputing (70), no. 13-15, pp. 2303-2323.

Weng, J.; McClelland, J.; Pentland, A.; Sporns, O.; Stockman, I.; Sur, M.; and E. Thelen (2001). Artificial intelligence. Autonomous mental development by robots and animals. Science 291, 599-600.

Wilson, F.R. (1998). The hand: How its use shapes the brain, language, and human culture. New York: Pantheon Books.