The Emergence of Self-Conscious Systems
From Symbolic AI to Embodied Robotics

Klaus Mainzer

Chair for Philosophy of Science, Institute of Interdisciplinary Informatics
University of Augsburg, D-86135 Augsburg, Germany
klaus.mainzer@phil.uni-augsburg.de

Abstract

Since hundred millions of years, the natural evolution on Earth has developed nervous systems with increasing complexity. They work according to algorithms of neurochemistry and equip organisms with self-adapting, self-controlling, and self-conscious features. But the laws of evolution could also admit completely different forms of life on different material basis – and perhaps they have emerged elsewhere in the universe. Therefore, humans and animals are only special cases of intelligent systems which have emerged on Earth under more or less random conditions. They are neither goals nor in the center of evolution. Traditional AI had tried to imitate the human mind by symbolic programming with only modest success. In a technical evolution of embodied robotics, artificial forms of life and self-conscious systems could emerge with new self-organizing features. But, like in natural evolution, self-organization does not automatically lead to desired results. Therefore, controlled emergence is a challenge of future technology. A new moral responsibility is demanded in order to handle human-robotic interaction which is evolving in a technical co-evolution.

Classical AI: Symbolic Representation & Control

Knowledge representation used today in database applications, artificial intelligence, software engineering, and many other disciplines of computer science has deep roots in logic and philosophy (Mainzer 2003). In the beginning, there was Aristotle (384-322 B.C.) who developed logic as a precise method for reasoning about knowledge. Syllogisms were introduced as formal patterns for representing special figures of logical deductions. According to Aristotle, the subject of ontology is the study of categories of things that exist or may exist in some domain.

In modern times, Descartes considered the human brain as a store of knowledge representation. Recognition was made possible by an isomorphic correspondence between internal geometrical representations (ideae) and external situations and events. Leibniz was deeply influenced by these traditions. In his "mathesis universalis", he required a universal formal language (lingua universalis) to represent human thinking by calculation procedures and to implement them to mechanical calculating machines. An "ars iudicandi" should allow every problem to be decided by an algorithm after representation in numeric symbols. An „ars inveniendi“ should enable users to seek and enumerate desired data and solutions of problems. In the age of mechanics, knowledge representation was reduced to mechanical calculation procedures.

In the 20th century , computational cognitivism arose on the background of Turing's theory of computability. In his functionalism, the hardware of a computer is related to the wetware of human brain. The mind is understood as the software of a computer. Turing argued: If human mind is computable, it can be represented by a Turing program (Church's thesis) which can be computed by a universal Turing machine, i.e. technically by a general purpose computer. Even if people do not believe in Turing's strong AI-thesis, they often claim classical computational cognitivism in the following sense: Computational processes operate on symbolic representations referring to situations in the outside world. These formal representations should obey Tarski's correspondence theory of truth: Imagine a real world situation X_1 (e.g., some boxes on a table) which is encoded by a symbolic representation A_1=encode(X_1) (e.g., a description of the boxes on the table). If the symbolic representation A_1 is decoded, then we get the real world situation X_1 as its meaning, i.e. decode(A_1)=X_1. A real-world operation T (e.g., a manipulation of the boxes on the table by hand) should produce the same real-world result A_2, whether performed in the real world or on the symbolic representation: decode(encode(T)(encode(X_1)))=T(X_1)=X_2. Thus, there is an isomorphism between the outside situation and its formal representation in Cartesian tradition. As the symbolic operations are completely determined by algorithms, the real-world processes are assumed to be completely controlled. Therefore, classical robotics operate with completely determined control mechanisms.

New AI: Self-Organization and Controlled Emergence

Knowledge representations with ontologies, categories, frames, and scripts of expert systems work along this line. But, they are restricted to a specialized knowledge base without the background knowledge of a human expert. Human experts do not rely on explicit (declarative) rule-

based representations, but on intuition and implicit (procedural) knowledge (Dreyfus 1979). Further on, as already Wittgenstein knew, our understanding depends on situations. The situatedness of representations is a severe problem of informatics. A robot, e.g., needs a complete symbolic representation of a situation which must be updated if the robot's position is changed. Imagine that it surrounds a table with a ball and a cup on it. A formal representation in a computer language may be ON(TABLE,BALL), ON(TABLE,CUP), BEHIND(CUP,BALL), etc. Depending on the robot's position relative to the arrangement, the cup is sometimes behind the ball or not. So, the formal representation BEHIND(CUP,BALL) must always be updated in changing positions. How can the robot prevent incomplete knowledge? How can it distinguish between reality and its relative perspective? Situated agents like human beings need no symbolic representations and updating. They look, talk, and interact bodily, e.g., by pointing to things. Even rational acting in sudden situations does not depend on internal representations and logical inferences, but on bodily interactions with a situation (e.g., looking, feeling, reacting).

Thus, we distinguish formal and embodied acting in games with more or less similarity to real life: Chess, e.g., is a formal game with complete representations, precisely defined states, board positions, and formal operations. Soccer is a nonformal game with skills depending on bodily interactions, without complete representations of situations and operations which are never exactly identical. According to Merleau-Ponty, intentional human skills do not need any internal representation, but they are trained, learnt, and embodied in an optimal „gestalt" which cannot be repeated (Merleau-Ponty 1962). An athlete like a pole-vaulter cannot repeat her successful jump like a machine generating the same product. Husserl's representational intentionality is replaced by embodied intentionality. The embodied mind is no mystery. Modern biology, neural, and cognitive science give many insights into its origin during the evolution of life.

The key-concept is self-organization of complex dynamical systems (Mainzer 2004, Mainzer 2005). The emergence of order and structures in nature can be explained by the dynamics and attractors of complex systems. They result from collective patterns of interacting elements in the sense of many-bodies problems that cannot be reduced to the features of single elements in a complex system. Nonlinear interactions in multicomponent („complex") systems often have synergetic effects, which can neither be traced back to single causes nor be forecasted in the long run or controlled in all details. The whole is more than the sum of its parts. This popular slogan for emergence is precisely correct in the sense of nonlinearity.

The mathematical formalism of complex dynamical systems is taken from statistical mechanics. If the external conditions of a system are changed by varying certain control parameters (e.g., temperature), the system may undergo a change in its macroscopic global states at some critical point. For instance, water as a complex system of molecules changes spontaneously from a liquid to a frozen state at a critical temperature of zero Celsius. In physics, those transformations of collective states are called phase transitions. Obviously they describe a change of self-organized behavior between the interacting elements of a complex system. The suitable macrovariables characterizing the change of global order are denoted as „order parameters". They can be determined by a linear-stability analysis (Mainzer 2005). From a methodological point of view, the introduction of order parameters for modeling self-organization and the emergence of new structures is a giant reduction of complexity. The study of, perhaps, billions of equations, characterizing the behavior of the elements on the microlevel, is replaced by some few equations of order parameters, characterizing the macrodynamics of the whole system. Complex dynamical systems and their phase transitions deliver a successful formalism to model self-organization and emergence. The formalism does not depend on special, for example, physical laws, but must be appropriately interpreted for different applications.

There is a precise relation between self-organization of nonlinear systems with continuous dynamics and discrete cellular automata. The dynamics of nonlinear systems is given by differential equations with continuous variables and a continuous parameter of time. Sometimes, difference equations with discrete time points are sufficient. If even the continuous variables are replaced by discrete (e.g., binary) variables, we get functional schemes of automata with functional arguments as inputs and functional values as outputs. There are classes of cellular automata modeling attractor behavior of nonlinear complex systems which is well-known from self-organizing processes.

But in many cases, there is no finite program, in order to forecast the development of future patterns. In general, there are three reasons for computational limits of systems dynamics: 1) A system may be undecidable in a strict logical sense. 2) Further on, a system can be deterministic, but nonlinear and chaotic.. In this case, the system depends sensitively on tiny changes of initial data in the sense of the butterfly effect. Long-term forecasting is restricted, and the computational costs of forecasting increase exponentially after some few steps of future predictions. 3) Finally, a system can be stochastic and nonlinear. In this case, only probabilistic predictions are possible. Thus, pattern emergence cannot be controlled in any case.

Self-organization and pattern emergence can also be observed in neural networks, working like brains with appropriate topologies and learning algorithms. A simple robot with diverse sensors (e.g., proximity, light, collision) and motor equipment can generate complex behavior by a self-organizing neural network. In the case of a collision

with an obstacle, the synaptic connections between the active nodes for proximity and collision layer are reinforced by Hebbian learning: A behavioral pattern emerges, in order to avoid collisions in future.

Obviously, self-organization leads to the emergence of new phenomena on sequential levels of evolution. Nature has demonstrated that self-organization is necessary, in order to manage the increasing complexity on these evolutionary levels. But nonlinear dynamics can also generate chaotic behavior which cannot be predicted and controlled in the long run. In complex dynamical systems of organisms monitoring and controlling are realized on hierarchical levels. Thus, we must study the nonlinear dynamics of these systems in experimental situations, in order to find appropriate order parameters and to prevent undesired emergent behavior as possible attractors. The challenge of complex dynamical systems is controlled emergence.

A key-application is the nonlinear dynamics of brains. Brains are neural systems which allow quick adaption to changing situations during the life-time of an organism. In short: They can learn. The human brain is a complex system of neurons self-organizing in macroscopic patterns by neurochemical interactions. Perceptions, emotions, thoughts, and consciousness correspond to these neural patterns. Motor knowledge, e.g., is learnt in an unknown environment and stored implicitly in the distribution of synaptic weights of the neural nets. In the human organism, walking is a complex bodily self-organization, largely without central control of brain and consciousness: It is driven by the dynamical pattern of a steady periodic motion, the attractor of the motor system. Motor intelligence emerges without internal symbolic representations.

What can we learn from nature? In unknown environments, a better strategy is to define a low-level ontology, introduce redundancy – and there is a lot in the sensory systems, for example – and leave room for self-organization. Low-level ontologies of robots only specify systems like the body, sensory systems, motor systems, and the interactions among their components, which may be mechanical, electrical, electromagnetic, thermal etc. According to the complex systems approach, the components are characterized by certain microstates generating the macrodynamics of the whole system.

Take a legged robot. Its legs have joints that can assume different angles, and various forces can be applied to them. Depending on the angles and the forces, the robot will be in different positions and behave in different ways. Further, the legs have connections to one another and to other elements. If a six-legged robot lifts one of the legs, this changes the forces on all the other legs instantaneously, even though no explicit connection needs to be specified (Pfeifer and Scheier 2001). The connections are implicit: They are enforced through the environment, because of the robot‘s weight, the stiffness of its body, and the surface on which it stands. Although these connections are elementary, they are not explicit and included if the designer wished. Connections may exist between elementary components that we do not even realize. Electronic components may interact via electromagnetic fields that the designer is not aware of. These connections may generate adaptive patterns of behavior with high fitness degrees (order parameter). But they can also lead to sudden instability and chaotic behavior. In our example, communication between the legs of a robot can be implicit. In general, much more is implicit in a low-level specification than in a high-level ontology. In restricted simulated agents with bounded knowledge representation, only what is made explicit exists, whereas in the complex real world, many forces exist and properties obtain, even if the designer does not explicitly represent them. Thus, we must study the nonlinear dynamics of these systems in experimental situations, in order to find appropriate order parameters and to prevent undesired emergent behavior as possible attractors.

But not only „low level“ motor intelligence, but also „high level“ cognition (e.g., categorization) can emerge from complex bodily interaction with an environment by sensory-motor coordination without internal symbolic representation. We call it „embodied cognition“: An infant learns to categorize objects and to build up concepts by touching, grasping, manipulating, feeling, tasting, hearing, and looking at things, and not by explicit representations. The categories are based on fuzzy patchworks of prototypes and may be improved and changed during life. We have an innate disposition to construct and apply conceptual schemes and tools (in the sense of Kant).

Moreover, cognitive states of persons depend on emotions. We recognize emotional expressions of human faces with pattern recognition of neural networks and react by generating appropriate facial expressions for non-verbal communication. Emotional states are generated in the limbic system of the brain which is connected with all sensory and motoric systems of the organism. All intentional actions start with an unconscious impulse in the limbic system which can be measured some fractals of a second before their performance. Thus, embodied intentionality is a measurable feature of the brain (Freeman 2004). Humans use feelings to help them navigate the ontological trees of their concepts and preferences, to make decisions in the face of increasing combinatorical complexity: Emotions help to reduce complexity.

The embodied mind (Balke and Mainzer 2005) is obviously a complex dynamical system acting and reacting in dynamically changing situations. The emergence of cognitive and emotional states is made possible by brain dynamics which can be modeled by neural networks. According to the principle of computational equivalence (Mainzer 2003, Mainzer 2004), any dynamical system can be simulated by an appropriate computational system. But, contrary to Turing’s AI-thesis, that does not mean

computability in any case. In complex dynamical systems, the rules of locally interacting elements (e.g., Hebb's rules of synaptic interaction) may be simple and programmed in a computer model. But their nonlinear dynamics can generate complex patterns and system states which cannot be forecast in the long run without increasing loss of computability and information. The main reason is the stochastic and nonlinear dynamics of the brain. There is a continuous random noise of firing neurons in the background of any measured neural signal. Further on, nonlinear stochastic dynamics can lead to stochastic chaos, depending sensitively on tiny changing conditions. Thus, artificial minds could have their own intentionality, cognitive and emotional states which cannot be forecast and computed like in the case of natural minds. Limitations of computability are characteristic features of complex systems.

In a dramatic step, the complex systems approach has been enlarged from neural networks to global computer networks like the World Wide Web. The internet can be considered as a complex open computer network of autonomous nodes (hosts, routers, gateways, etc.), self-organizing without central mechanisms. Routers are nodes of the network determining the local path of each information packet by using local routing tables with cost metrics for neighboring routers. These buffering and resending activities of routers can cause congestions in the internet. Congested buffers behave in surprising analogy to infected people. There are nonlinear mathematical models describing true epidemic processes like malaria extension as well as the dynamics of routers. Computer networks are computational ecologies (Mainzer 2004).

But complexity of global networking not only means increasing numbers of PCs, workstations, servers, and supercomputers interacting via data traffic in the internet. Below the complexity of a PC, low-power, cheap, and smart devices are distributed in the intelligent environments of our everyday world. Like GPS in car traffic, things in everyday life could interact telematically by sensors. The real power of the concept does not come from any one of these single devices. In the sense of complex systems, the power emerges from the collective interaction of all of them. For instance, the optimal use of energy could be considered as a macroscopic order parameter of a household realized by the self-organizing use of different household goods according to less consumption of electricity during special time-periods with cheap prices. The processors, chips, and displays of these smart devices don't need a user interface like a mouse, windows, or keyboards, but just a pleasant and effective place to get things done. Wireless computing devices on small scales become more and more invisible to the user. Ubiquitous computing enables people to live, work, use, and enjoy things directly without being aware of their computing devices.

Self-Conscious Systems, Human Responsibility, and Freedom

Obviously, interacting embodied minds and embodied robots generate embodied superorganisms of self-organizing information and communication systems. What are the implications of self-organizing human-robot interaction (HRI)? Self-organization means more freedom, but also more responsibility. Controlled emergence must be guaranteed in order to prevent undesired side-effects. But, in a complex dynamical world, decision-making and acting is only possible under conditions of bounded rationality. Bounded rationality results from limitations on our knowledge, cognitive capabilities, and time. Our perceptions are selective, our knowledge of the real world is incomplete, our mental models are simplified, our powers of deduction and inference are weak and fallible. Emotional and subconscious factors affect our behavior. Deliberation takes time and we must often make decisions before we are ready. Thus, knowledge representation must not be restricted to explicit declarations. Tacit background knowledge, change of emotional states, personal attitudes, and situations with increasing complexity are challenges of modeling information and communication systems. Human-oriented information services must be improved in order to support a sustainable information world.

While a computational process of, e.g., a PC is running, we often know neither the quality of the processing or how close the current processing is to a desired objective. Computational processes seldom have intermediate results to tell us how near is the current process to any desired behavior. In biological systems, e.g., we humans experience in a sense that we know the answer. In a kind of recursive self-monitoring, some internal processes observe something about our cognitive states that help us to evaluate our progress. The evolutionary selection value of self-reflection is obvious: If we have these types of observations available to us, we can alter our current strategies according to changing goals and situations. Engineered systems have some counterparts to kinesthetic feedbacks one finds in biological systems. But the challenge is to create feedback that is meaningful for decisions in a system that can reflectively reason about its own computations, resource use, goals and behavior within its environment. This kind of cognitive instrumentation of engineered systems (Bellman 2005) can only be the result of an artificial evolution, because cognitive processes of humans with their multiple feedback processes could also only develop during a long history of evolution and individual learning.

Thus, we need generative processes, cognitive instrumentation, and reflective processes of systems in order to handle the complexity of human-robot interactions. Biological systems take advantage of layers

with recursive processing of self-monitoring and self-controlling from the molecular and cellular to the organic levels. Self-reflection leads to a knowledge that is used by a system to control its own processes and behaviors. What distinguished self-reflection from any executive control process is that this reflection involves reasoning about that system, being able to determine or adjust its own goals because of this reflection. But self-reflection must be distinguished from self-consciousness. Consciousness is at least partly about the feeling and experience of being aware of one's own self. Therefore, we could construct self-reflecting systems without consciousness that may be better than biological systems for certain applications. It is well known that technical instruments (e.g., sensors) already surpass the corresponding capacities of natural organisms with many orders of magnitude. Self-reflecting systems could help to improve self-organization and controlled emergence in a complex world.

But, how far should we go? Self-consciousness and feeling are states of brain dynamics which could, at least in principle, be simulated by computational systems. The brain does not only observe, map, and monitor the external world, but also internal states of the organism, especially its emotional states. Feeling means self-awareness of one's emotional states. In neuromedicine, the "Theory of Mind" (ToM) even analyzes the neural correlates of social feeling which are situated in special areas of the neocortex. People, e.g., suffering from Alzheimer disease, loose their feeling of empathy and social responsibility because the correlated neural areas are destroyed. Therefore, our moral reasoning and deciding have a clear basis in brain dynamics which, in principle, could be simulated by self-conscious artifical systems. In highly industrialized nations with advanced aging, feeling robots with empathy may be a future perspective for nursing old people when the number of young people engaged in public social welfare decrease and the personal costs increase dramatically.

Humans are not in the center of the universe and evolution, but they are in the center of their history and culture. The concept of human personality refers to human historicity, self-identity, intentionality, and embedding in the intimacy of human social and cultural identity. Therefore, AI, cognitive science, and computer science have to take care of humans as a value and purpose in their own (Kant: "self-purpose") which should be the measure of human technology. That is a postulate of practical reason which has developed and approved itself in evolution and history of mankind. In principle, future technical evolution could generate self-conscious systems which are not only human copies ("clones"), but artificial organisms with their own identity and intimacy which would differ from ours. But why should we initiate an evolution separating from human interests? AI, bio-, information and communication technology should be developed as human service in order to heal and help in the tradition of medicine.

This is a humanistic vision different from sciencefiction dreams which only trust in the technical dynamics of increasing computational capacity, leading automatically to eternal happiness and computational immortality. In nonlinear dynamics, there is no guarantee for final stable states of order. We need "order parameters" for moral orientation in changing situations of our development. We should not trust in the proper dynamics of evolution, and we should not accept our deficient nature which is more or less the result of a biological random game and compromise under changing conditions on the Earth. The dignity of humans demands to interfere, change, and improve their future. But, it should be our decision who we want to be in future, and which kind of artificial intelligence and artificial life we need and want to accept besides us.

References

Balke, W.-T., and Mainzer, K. 2005. Knowledge Representation and the Embodied Mind: Towards a Philosophy and Technology of Personalized Informatics. In *Lecture Notes of Artificial Intelligence* 3782: *Professional Knowledge Management*, 586-597. Berlin: Springer

Bellman, K.L. 2005. Self-Conscious Modeling. *it – Information Technology* 4:188-194

Dennett, C.D. 1998. *Brainchildren: Essays on Designing Minds*. Cambridge Mass.: MIT Press

Dreyfus, H.L. 1979. *What Computer's can't do – The Limits of Artificial Intelligence*. New York: Harper & Row

Dreyfus, H.L. 1982. *Husserl, Intentionality, and Cognitive Science*. Cambridge Mass.: MIT Press

Floridi, L. ed. 2004. *Philosophy of Computing and Information*. Oxford: Blackwell

Freeman, W.J. 2004. How and why Brains create Sensory Information. *International Journal of Bifurcation and Chaos* 14:515-530

Mainzer, K. 2003. *KI - Künstliche Intelligenz. Grundlagen intelligenter Systeme*. Darmstadt: Wissenschaftliche Buchgesellschaft

Mainzer, K. 2004. *Thinking in Complexity. The Computational Dynamics of Matter, Mind, and Mankind.* New York : 4th enlarged edition Springer

Mainzer, K. 2005. *Symmetry and Complexity. The Spirit and Beauty of Nonlinear Science*. Singapore: World Scientific

Merleau-Ponty, M. 1962. *Phenomenology of Perception*. Routledge & Kegan Paul

Pfeifer, R., and Scheier C. 2001. *Understanding Intelligence*. Cambridge Mass.: MIT Press

Searle, J.R. 1983. *Intentionality. An Essay in the Philosophy of Mind*. Cambridge University Press