

# The Natural History of Agents

George Kampis  
Dept. of History and Philosophy of Science  
Eötvös University, Budapest  
[gk@hps.elte.hu](mailto:gk@hps.elte.hu)

## Introduction

Writing the natural history of agents poses a double challenge. Not only is it unclear (despite recent interest in agent based software technology) what should we understand under the term 'agent'. It is also a question, what natural history has to offer to the study of agents.

One property the software objects called 'agents' have in common is that they show (are expected to show, believed to show, promised to show in return for the next budget...) a certain kind of *intelligence*. "An agent is a software thing that knows how to do things that you could probably do yourself if you had the time", says Ted Selker of the IBM Almaden Research Centre (quoted from Janca 1995). They are like us, and we are like them. So is an agent a little software person, and if yes/no, how can we tell?

In this paper we briefly review the psychological and philosophical origins of some key concepts of "intelligence" which contribute to what is called agent thinking today. I will focus on ideas and characters of the XIX. century that turned out to be of key importance later, like William James, Franz Brentano, or intentionality.

## Exposition

How do we get there? The perhaps most obvious aspect of natural history is that of chronology, in other words, event history parametrized by time. If we took Nature's development and took a chronological picture of the natural history of agents, we could expect a report on fossile and more recent organisms, depicted in snapshots taken from the phylogeny of intelligence. Ultimately, we would end up discussing stuffed animals. Is the shrew-mouse more intelligent, or more agent-like, or whatever, than the nearby dinosaur? Look at those clever little fingers! Look at that oversize brain! And above all, look at those expressive eyes! Sure he is intelligent. Or is he not? Well, my kids don't stand stuffed animals. What a miss.

Here I will follow a different meaning of natural history. Natural history can also be understood more theoretically, as a recapitulation of the evolution of the functional principles or solution techniques as applied by Nature. Rather than a study in comparative zoology, psychology (robotology... and so on...), the natural history of agents becomes subordinate, under this view, to the philosophical question of what intelligence is about, and how it is manifested in various natural and possibly artificial forms. In this spirit, in the present paper we will talk about the first principles of animal and human intelligence, to shed light on what

makes them and us, animals and humans, so special - as opposed to a piece of stone or a soup of chemicals of which we all evolved.

As we will see, the essentials of the difference (if there is any) between agents and non-agents have to do with considerations of nineteenth century philosophical psychology, and that is the point where we find the birth of certain now-perennial problems of cognition, philosophy, and artificial intelligence. The basic ideas and indeed the entire problem set which was bootstrapped by these old and often underestimated authors are invariably present in today's cognitive science and philosophy of mind - so we have no other choice but to talk about the principles of agents from this double perspective, that of old and modern, history of science and philosophy of mind alike.

The justification of the foregoing historical remarks, if one needs a justification at all, lies in my personal skepticism with the rapidly changing fashion words. Concepts and fashions may change but the truly deep questions remain. There is a nice little story about the transitory and the permanent. One day a pilgrim came to the palace of the grand mogul, and praised the building as a "nice caravanserai". The mogul asked, angrily, for an explanation. Tell me then, who lived in this place before you?, asked the pilgrim. My father, was the response. And before him? His father, then his grandfather, his grand-grandfather, and so on, said the mogul impatiently. Well, said the pilgrim, what is a place where people come and go, if not a caravanserai?

## What is an Agent?

And now up to the work. How come that the agent idea became so powerful? Why do we find it useful, in the first place, to talk about programs as if they were people or animals? And why do we tend to personify objects at all? We talk to our car, we hate our failing computer, we praise a favorite book. The tremendous success of the agent metaphor suggests that we indeed find it easy and enlightening to talk about, design, and analyse things as if they were animate. But we do not personify everything and not always. Objects dealt with in this privileged way are typically man-made objects, and typically, machines. What are now the common properties of these systems, which are amplified when talking of them as agents?

Let us start our discussion by looking at some agent definitions in the trade. There is a wide variety of definitions, ranging from the online Software Agents Mailing List FAQ (Belgrave 1998, which says, laconically, that an agent is "An entity authorized to act on another's behalf") to such often-quoted works as (Wooldridge and Jennings 1995, Franklin and Graesser 1996), which are dedicated entirely to this problem and its aura.

Here is a short list of a few somewhat *ad hoc* properties that alone or together may be taken to define an agent:

- proactivity, the ability to effect actions that achieve their goals by taking the initiative;
- intentionality, the attribution of purposes, beliefs and desires;
- autonomy, the ability to operate independently of and unaided by a user;

I will first discuss the path from proactivity to intentionality, and I will discuss why many people believe there is little difference between the two. Then I will present some varieties of

intentionality, a notorious subject, which is often believed to separate science from the humanities. Finally, we will have some remarks concerning autonomy.

## From Proactivity to Intentionality

Proactivity is very simple as a *prima facie* concept. It begins where a behaving subject takes command of its own actions.

This is not so simple at a closer look. What is the behaving subject, that can take control of an action (or anything)? What is the mind? Is there a mind at all? Behaviorism, an early twentieth-century movement shaped by the physics-motivated norms of the "positive" empirical sciences, denied the existence and/or the importance of the mind altogether. It dismissed the mind as something which is non-empirical and consequently does not exist or should not be dealt with (the exact behaviorist position depends on the exact variety of behaviorism taken). Above all, behaviorism was a profound *reactive* paradigm. True believers of behaviorism were hoping to explain even the most contemplative human actions like the production of speech and text as responses to environmental stimuli.

Behaviorism so defined, proactivity is best imagined as the complete opposite of that. Instead of reply to input, an internal driving force, a spontaneous initiative that gives rise to behavior is postulated. Instead of the behavior's dependence on the content of external stimulus, we assume dependence on *a priori* and stimulus-independent organizing forms. Take the example of language. The behaviorist idea would be that language capacity is learned and therefore everybody capable of learning can also acquire language. The opposite idea is that language is a structural capability of the mind and you either have it or not by birth. The fact that animals cannot "really" learn language, despite all the results with Washoe, Koko, and others up to the most recent bonobos, is a score for the anti-behaviorists. Learning is far from being universal; animals may learn but they do not learn to use language.

Proactivity can be better perhaps understood if we start to examine its previous career in the last century. William James (1842-1910) will be the first name to mention. He was the maybe first genuinely American philosopher and certainly the most often quoted ever since. William James is generally praised for being the co-founder of the philosophical movement of pragmatism. For the purposes of the present paper another, less frequently mentioned aspect of James will be important. He believed in the freedom of the will. Even this alone would not make him very original; there were others like Kierkegaard before him. What makes him original is that in the context of the free will he spoke of a certain process of "self-making", of a "Promethean" character of the self, which invokes some internal power which is determined neither by nature nor by nurture (ie. neither by inheritance nor by social context). Accordingly, James spoke of the importance of the individual, and he considered the individual's consciousness as something that stands above the chain of actual events and happenings of life. Experiences do not just *happen*, he claimed, but are results of the individual's singular decisions and acts. There is a certain unique faculty in man, that makes all this possible. This faculty is the one that transforms the individual and, by reaching out, its environment.

Huh - yes, but how can this kind of proactivity be achieved? Can we isolate this faculty and build it into a machine? Let us be fair to James: from our present-day perspective he was a humble phenomenalist, and not an engineer, a naturalist or some kind of traditional

ontological philosopher. To put it differently: he was only interested in the appearances of the self, and not in its material origins. He asked only the *how* and not the *why* questions. No causal explanation was sought. To understand this platform, we should also understand that this is an age where philosophy and psychology did not separate yet and therefore psychology did not exist as a science of its own. Indeed it was James together with Franz Brentano (who we will meet soon), who made the first steps towards psychology as a discipline. James' observations and speculations about the self were important contributions to the birth of philosophy even if they are useless from an engineering perspective of science. There is also a deep personal element in James' story. When he was young, James was depressed and at one point almost committed suicide, but succeeded to return and to recover. His notions of (phenomenal) will and moral responsibility thus rested on the solid ground of his own life experience.

Another early reference is Henry Bergson (1859-1941), a student of the evolutionist Herbert Spencer. In his several works Bergson elaborated the theory of *élan vital*, or "life impulse". This is a much misunderstood theory. The popular accounts of *élan vital* depict the concept as unscientific, and perhaps anti-scientific, a close relative to old-fashioned vitalism, which maintained that living matter could only be produced by more living matter, a claim so splendidly refuted by the chemist Koehler. Popular interpretation places *élan vital* into the same pot with witchcraft and superstition. It was Bergson's refusal of contemporary biology that earned him this undeserved bad reputation. In actuality Bergson spoke of something quite interesting and elaborate. When talking about the impulse of life Bergson speaks about life as a productive phenomenon, as a process, a temporal unfolding, a vivid history. Bergson's dynamic self is both result and starting point of mental transformations. He criticizes the prevailing studies of inert structures and speaks of the self as a temporal unity. How new and how relevant this is, is best seen in a recent rediscovery of the temporal problems of cognition (van Gelder and Port 1996). For example, Bergson was among the first to raise the problem of personal identity. This problem consists in the difficulty of talking about a permanent single person (the fixed point of the "I") in the philosophical, scientific (and legal!) sense, against a background where everything that constitutes the mind may change over time. The same problem and essentially the same solution (that when talking about the identity of a person we should talk about the identity of a process that makes up this person, rather than about a single state of a process) reemerges in Dennett and MacIntyre in their recent theory of narrative identity.

Let us return to behaviorism and agent theories. From the philosophical point of view, reactivity is an empiricist idea, where the outer facts and the experience act together to determine mental content. In the same old spirit, the idea of proactivity leads to rationalism, or the primacy of the mind at its discretion.

Behaviorism dominated the first half of the century, mostly because it was a lot more modern and attractive for the scientifically trained than any other alternative. What could be wrong with doing science on the basis of observable facts and perhaps logic, as did the behaviorists? One category of observable facts concerning the mind's workings can be conceived as the mind's inputs, constituting what philosophers like Bertrand Russell called "sense data". For a while it looked as if sense data theories could explain behavior, which was the other observable thing, after all. So you needed no mind in between. All you needed was reflexes, conditioning, and stimulus-response relations. Yet there were some notable failures, such as the scandal of consciousness. If we have no mind how can we be conscious? The scientist can talk me into believing we don't have a mind but I *feel directly* that I have. The misfortune of

psychology is that what are called "third person" (hard access) and "first person" (privileged access) perspectives are intertwined there: it is easy for the researcher to dismiss other people's mind, but what will be with our own? And what about our dreams, plans, imagination, purpose, and other invisible actions of the mind? Consciousness may be a red herring for some people but we do have plans, don't we?

The change of the viewpoint arrived in the works of Karl Lashley, Noam Chomsky and others near the end of the 40's, and led to the two cognitive revolutions - cognitive psychology first, cognitive science second (Gardner 1985). Nothing could be more different from the behaviorist program than the cognitivist movement. Where behaviorists imagined an input-output machine, cognitivists spoke of internal states. Where behaviorists claimed behavioral competence by learning, cognitivists spoke of innate origins. Where behaviorists spoke of resulting behavior, the cognitivists spoke of preemptive structure.

There is a twist in the story, however, as every cognitive scientist knows. The twist has various forms; in one of its versions it has to do with a famous problem of the Turing test. In a citation classic from 1950, Alan M. Turing introduced the idea that a computer may simulate human thought. At a closer look, this is a combination of two ideas: that a computer program can stand for a mind, and that a proper proof of this pudding is the behavioral or simulation test proposed in the paper. Now despite several denials, modern artificial intelligence and cognitive science relies on a platform of residual methodological behaviorism, implied by Turing's suggestion. The residual behaviorism of cognitive science is also associated with the name of Hilary Putnam, whose notion of machine functionalism is another well-known concept in the search for software minds. Machine functionalism assumes that the mind is essentially like an algorithm. Algorithms have, among other things, the nice property of substrate independence, which means that we can dance or sing them as well as carrying them out on electronic computers, there will be no difference. Computations could be carried out on wooden sticks, for instance, and still we gained nothing but speed, as compared with a Pentium II PC. Anything can be a computer. In fact the situation is worse than that. Even within a single medium, say, the world of wooden sticks, there are infinitely many different ways of realizing any given algorithm. As a result of this structural indifference property, the performance of an algorithm can only be evaluated on the basis of, well - its behavior.

And there is another enigma - cognitivists do not use the word proactivity. Intelligence, knowledge, planning, internal states, even goals and desires are lightheartedly attributed to humans and computer programs aimed at simulating human mental machinery. But what would be like a computer program that behaves proactively? Agent technology may now try to contribute to the solution of this question, but one has to be careful: perhaps what agent technology emphasizes is just pre-wired structure and nothing more. Software agents may do things but so do other programs. Of course every program has certain pre-wired characters (in fact this is the essence of being a program: a program is a program only because and insofar as it has a well-defined effective procedure that it carries out with a certain degree of accuracy and with a production guarantee). Besides, most programs can do different things depending on the presence different inputs. In this sense any program is both reactive and proactive. (Franklin and Graesser 1996 asks: "Autonomous agents? What do you mean by that? A brief explanation is then followed by: But agents sound just like computer programs.") We immediately see this issue when pressing the keys of the keyboard. Key pressing has exactly the said kind of double nature.

It is the keyboard where the computer is controlled from, and we can do anything from there that is doable in the universe of effective procedures. We are the masters, and the machine is the slave, etc. On the other hand, it is the computer's internal design that determines the kind of interactions that the machine can take as inputs. If, for instance, there are no separate keys for the Hungarian long umlauts, and you do need them, trouble begins. If something is not on your roll-down menu, you can't click on it. Now it's the computer who controls you, now it's the other way around. The case for software agents is no exception. Are they now all proactive because they do things by themselves? Or are they just old input-output systems, like black boxes that carry out a certain transformation function, or like a slot-machine-like gadget that responds to a software environment - essentially in the same form as envisioned by the behaviorists? Is there a difference between these options? What, exactly, would be the difference? And if there is one, does it matter?, given that programs are what they are - codes for algorithms.

As far as philosophy is the enterprise of turning the obvious into the unbelievable, this will be a fine philosophical topic for us to dwell on further, if in a somewhat different form.

## **Varieties of Intentionality**

Both early conceptions, those of James and Bergson, placed some agency into the mind that makes it possible for the subject to "reach out" towards the world. There is some supposed directedness. It is perhaps no accident that "reaching out", "pointing to", "being in a referential state towards" or simply "aboutness" plays a central role in a very important tradition that hopes to delineate the human world and the human sciences from the rest.

"Aboutness" enters modern philosophy via the notion of intentionality. In a broad sense, intentionality amounts to the attribution of mental states: hopes, fears, plans, and, above all, beliefs and desires (i.e. knowledge and purpose, in lay terms, if you wish). Every treatment of intentionality hastens to warn the reader that the technical meaning of the word does not come from the ordinary sense of intention or purpose, so be warned. The modern career of the concept begins with Franz Brentano (1838-1917). The roots are much older, as always.

Intentionality goes back to the medieval scholastics and to the study of nonexistent objects such as unicorns. To think of a horse is to maybe to use direct percepts and their memory traces, but how can we think of an angel or a unicorn? (The tempting suggestion, don't think of them, does not work; most people believe there is something wrong with the sentence "Unicorns have two horns" while the same people readily accept "Krgzxx have two horns", perhaps because the Krg...stuff is something we find it easier not to think about or to imagine after all.) The monastic philosophy of the scholastics introduced intentionality, among other things, to name (and by the act of naming, to kind of explain) the human faculty which enables us to think of things that are absent.

The radical Brentano thesis is that internal objects (called intentional objects) exist in the mind, and that this distinguishes humans from the rest of the Universe. In other words, Brentano put forward a demarcation thesis which functions as a licence for psychology to work as a separate discipline. Besides, the thesis invokes an own methodology. It says that there are the natural sciences, on the one hand, which deal with non-intentional objects, and there is psychology, which deals with the intentional ones - minds.

Let us stop for a while to contemplate the feasibility of the conception. It is true that thought always has a target, an object, or a referent. It is not possible to just sit and think without thinking *about something* (one of the problems parents face when telling kids to "think". Think, fine, but think about what?). The Brentano thesis amplifies the outward-boundness of thought by positing an irreducibility for psychological experience, and by claiming that every experience involves intentional objects. Intentionality taken at face value amounts to a conception of "immanent intentionality". Following Aristotelian and medieval roots immanent intentionality assumes that the intentional object is literally present in the experience (or in the experiencer's mental states). Thought is a special kind of matter which unlike "normal" matter refers to other matter by meaning.

The demarcatory nature of intentionality should be taken quite seriously. There is a powerful tradition in European thought which proceeds through students of Brentano and their students to infamously anti-scientistic philosophic conceptions: we have a line of development from A. Meinong, E. Husserl, and M. Heidegger to M. Merleau-Ponty and his existential phenomenology, which takes the position that "life is nothing but meaning". Existential phenomenology pretends that not sensory appearances, sense data, and the usual physically existing components of the external world are the primary constituents of our Universe, but meanings and essences.

There is, at the same time, a more critical development concerning intentionality, where R. Chisholm, W. Sellars, W.v.O. Quine, and most recently, D.C. Dennett use the concept in a framework compatible with natural science, at least to some (various) degree(s).

This can be approached via derived intentionality. Whereas beliefs and desires as attributes of mental states were thought of by Brentano and friends as *immanent* features, in ordinary discourse intentional notions are typically used in a more permissive sense. For example, maps, pictures and books, or the famous example of the light switch in Searle (1980) show a certain kind of intentionality in the sense that these objects convey and express meaning, belief and purpose. The map shows, the picture depicts, the book tells. To be sure, most people are ready to agree that the said artifacts *derive* their intentionality, and that they do so by their social use. In other words, a book as a physical object is rather dull, but becomes vivid and evokes definitive thoughts when read. That is, books are really *about* things, but this is only true against a default backdrop which contains a default reader (who speaks the same language as the book, and so on. To qualify as a reader we have to fulfill a number of conditions). A light switch is a similarly dull electric device but has function and purpose - at least for those who know how to use it.

Intentionality and the attribution of mental states are closely related to another topic, folk psychology. Folk psychology deals with belief-desire accounts of human behavior. The name is self-explanatory, that is how the folk (i.e. us, in the everyday context) do psychology. The average person gives an account of other people's behavior by assuming or positing certain non-observable structures. "Joe went to the bar to buy a beer to relieve his thirst". Implied in such a description there are several assumptions. To say or understand this sentence we must take it as given that Joe knows, or believes that beer relieves thirst, that beer is sold at the bar, that going to the bar is a way of getting beer, and so on. Also we assume without any further study that if Joe is thirsty he wants (i.e. desires) a beer or some other drink, and that this why he goes to the bar and buys beer. How intricate this web of preliminary assumptions can be is best seen on the efforts of the artificial intelligencers to make computer programs "understand" even the simplest dialogs which involve accounts of genuinely human action.

To avoid philosophical difficulties which are complicated enough for another tutorial paper, let us only require by "understanding" that the machine gives a proper response (any response acceptable for a human) in the sense of the already criticized Turing test. The frustrating results of recent attempts can be seen from the transcripts of each year's Loebner competition (see e.g. <http://hps.elte.hu/~gk/loebner.html>).

An interesting feature of folk psychology is that it is something difficult to be without. Behaviorists volunteered to deprive themselves of the convenient language of meaningful accounts. Folk psychology restores meaning. The methodologically consistent behaviorist must confine himself to a mere physical description of the animal's moving body parts, without any mention of the meaning that these motions together might serve. But that is not how real field observations of behavior are made. Even the behaviorist uses shorthands. Under "normal" circumstances (that is, when we are not posing ourselves as behaviorists) we "see" the animal drink water, for instance. In fact what we can see in the physiological sense is that the animal moves its legs and mouth in a specific way, and that the water level decreases in the drinking fountain. The rest is interpretation. We simply add further elements to make sense. Also sense, the concept of drinking, simplifies the story. Drinking is a functional rather than physical concept, an animal might drink differently every time and still will be recognized as drinking. Similar interpretations of still higher levels make it possible for the scientist to perform experiments and to make observations about more complicated behaviors. Recent cognitive ethology has proved that semi-naturally held chimpanzee groups show intricate social behaviors that include power fights and co-operation among animals ("Chimpanzee Politics", de Waal 1982). You already guessed it, power fights are not directly observable either. So what is a power fight? It is that animals perform complex actions which cannot be consistently described by the scientist in any meaningful way if not called power fights and plans thereof. A young male develops an alliance (itself an interpretation, to be justified separately in the light of the rest) in order to win a position months later. It seems this kind of behavior is entirely inaccessible for science unless in the light of the power plan whose only confirmation is its consummation. That is, without the additive of folk psychology we are entirely deprived of the goodies of social ethology.

But there are problems. One of the troubles with folk psychology lies in the fact that it is difficult to tell what folk psychology actually is. Is it a commonsense theory underlying everyday behavior? The practice of ascribing mental states? Or a conceptual cornerstone of cognitive science, which on wholesale was developed in mentalistic terms? All of these? Can different meanings be kept separately? Strategies concerning folk psychology vary depending on the answer to these (and other) questions.

Just a few possibilities concerning folk psychology:

Eliminativism (Churchland, Stich) maintains that folk psychology is an empirical theory - and that it is a bad one. In other words, eliminativists assume that in order to be consistent folk psychology must pretend to be a scientific theory of a naïve kind, much like alchemy or the phlogiston theory of burning, from the present-day perspective of chemistry and oxygen. The message is that a scientific account of behavior must eliminate romantic old ideas by studying the neural structures that make it possible for the brain to perform a given behavior.

Mixed pragmatic positions. Accept that folk psychology is based on empirical generalizations but deny that these are of scientific nature. The concept of a warm coat is something we develop empirically (there is mom's fur coat, our own feathercoat, etc.) but this concept does

not refer to any clearly individuated scientific category. That is, there are many words that make sense but are not scientific.

Folk psychological realism. Another pragmatic attitude according to which the chimp story will never get any better. Like with engineering where the Maxwell equations may completely describe a computer but they are useless for that purpose. The engineer will never use electrodynamics because there is no upward translation, so knowing the electrodynamics gives in itself no knowledge of the electronics. Engineering will not develop by developing better low level descriptions but by applying the known higher level tools.

Naturalism (Dennett, Millikan). It is an adaptive evolutionary strategy for an organism to interpret other animals' behavior in terms of its beliefs and desires. That is the same thing as what we humans do. If I want to survive, it is perhaps justified but not very handy to study the lion's motion in clumsy behaviorist terms (object #17 raised right paw). By assuming, on the other hand, that the movement has a meaning and that the movement's meaning is that the lion wants to come, get and eat me, I get a ready-made model for the fast prediction of the lion's behavior. "Prediction for survival" is the key concept here. The attribution of mental states appears to serve this purpose very well. Our life experience suggests that we are indeed normally able to anticipate other people's (and perhaps other animals') reactions on the basis of their suspected beliefs, desires and plans, and to shape our own preventive behaviors in response.

Naturalistic versions of intentionality permit a further trick, best seen in Dennett. The magic is a substitution of derived intentionality (that of books) for inherent intentionality (that of Brentano's psychological subjects). Dennett is often described as an instrumentalist philosopher. From the tenet of instrumentalism it follows that all we can have in science is some simplified posit, a working model of reality, that is, a description serving a purpose. Within this framework it is not meaningful, therefore, to ask, whether there really exist an immanent intentional object in the mind. Because, if the answer is yes, this leads to a self-refuting claim: by invoking immanent intentionality under the instrumentalist position, we in fact attribute a model, a scientific description to the object of study - thereby ending up with the opposite, viz. derived intentionality, "in the eye of the observer".

Finally, let us say a straight word about how things stand on intentionality. From the standpoint of natural science, intentionality cannot be additive to the already known and yet to be understood material properties. It is cheap to say therefore, that intentionality is a pseudo-problem. On the other hand, many theoreticians accept that minds *are* different from stones, just we do not know how. Whether programs and software agents are more like stones or more like minds, is an open question. There is no definitive solution to any of these problems but dealing with them probably clarifies what is at stake.

## **Autonomy**

And there is, finally, the condition of autonomy to discuss. Problems and criteria of autonomy come from robotics but can be applied to software agents as well. Autonomy or self-sufficiency enters straightforwardly as a natural requirement for systems that are to perform a permanent or recurrent task without further control or intervention.

There are trivial blends of autonomy not worth wanting or reflecting on. In the world of mobile robots autonomy began with dropping the connecting cables between robot arm and mainframe computer. This happened at a cross-point of hardware technology where it became possible to mount both batteries and processor on piggyback. Higher versions of autonomy are more elegant and unlike the heavy batteries make life easier. These more elegant versions of autonomy include monitoring of the internal state, self-care (such as the regular loading of batteries, cleaning of mechanical parts, or updating of software), partial or total self-repair, and even adaptivity to changing environments so as to retain operability.

The history of autonomy goes back to the early days of cybernetics and system theory when L. von Bertalanffy, W.R. Ashby, D. MacKay and others have pioneered the notions of homeostasis, dynamical equilibrium, and feedback control. Some of these concepts might cling obsolete today, especially feedback control, which became a standard part of engineering. In their original context these were bold and rich concepts, however, ones that may help us understand some of the crucial problems around the notion of autonomy. Of them, homeostasis is perhaps the most complex and most difficult to formulate. Homeostasis is most often understood as the ability to recover and maintain function in a wide range of perturbations, by recurring to an actively maintained internal balance. Hormonal control in organisms, or the diverting reactions of the immune system are examples. Beyond these, homeostasis is about the organization of the living which makes it possible for the control mechanisms to appear and work in the first place.

Many people found it difficult to understand this aspect of homeostasis. Once the proper molecules are there, these molecules exert the biochemical control they way they do, and there is no need for any further "organization". Indeed there is no room for anything else then the molecules, because the whole organism consists of nothing but molecules. Or? This is a crucial point for I believe here lies one of the keys to the discussions on autonomy. Compare a living organism with a car. Typical cars unlike those in Baywatch tend not to react to their environment in any interesting way. There are two types of interactions with a car from the outside: those which leave the car unchanged (like leaning against it) and those that cause damage. And by damage we mean damage - in lack of self-repair or a diverting mechanism every change is permanent. Cells are different. The cell can "undo" things, and in fact unless you come with a hammer it undoes almost everything, be it a mechanical, chemical, or or electrical kind of change (did we leave out something?). Homeostasis now emphasizes the importance of this difference, pointing out that different classes of systems exist, ones which are built such that change can be prevented or compensated for, and others, where this is not the case. More modern concepts such as "resilience" in ecology and "graceful degradation" (Clark 1992) in artificial neural networks (ANN-s) obfuscate this difference while drawing on the same ground idea. If we cut a tree the forest survives. If we catch a bird or pick a flower, not very much will change either. If we cut all trees, catch all birds, pick all flowers, nothing remains. It can't be any other way. But the interesting thing is that the curve between none and all is not linear. Damages up to a certain point will be completely undone; not much above this point the system dies. Where lies the limit of internal recovery and what is its mechanism, are questions that refer to the resilience problem of ecosystems. Is resilience not a surprising property? From the organizational perspective, the big discovery is that there is resilience at all (compare this with the chaos theory fantasies, the devastating effects of the flapping wing of a butterfly, and the like). That is what homeostasis is about. The ANN story is a bit different, the ANN phenomenon is a consequence of distributed representation: cut off the half of an ANN "brain" and no information will be lost, or not completely. In exchange, all information pieces will be influenced to some, usually differing, extent.

The organizational element in control has drawn the attention of several theoreticians. A first elaboration is due to N. Rashevsky, one of the founders of mathematical biology. In his relational biology (which was later developed by R. Rosen into the theory of metabolism-repair systems) Rashevsky introduced mathematical models for the complex interaction between the organism as a functional whole and its parts in the interest of survival. This can be thought of as the mathematics of systems that renew themselves. Much as intentionalists, Rashevsky and Rosen did not deal with causal questions, however. They were only interested in the mathematical description of the renewal process *once it is there*.

But is self-maintenance and self-repair real? Are they theoretically possible at all? For example, is it possible for a system to describe itself in order to use this description as a master plan for later repairs? This is a famous question, which was first considered, as far as one can tell, by John von Neumann while working on his self-reproducing cellular automata (von Neumann 1966). Von Neumann suspected a paradox here: while describing itself, the machine would need to change its states, thereby invalidating the description of these same states. R. Laing clarified the logical nature of self-description in 1977 and showed, constructing an example, that self-describing machines are indeed possible.

A fresh and according to many accounts controversial look at the problems of autonomy and self-maintenance is the theory of autopoiesis (note the extra i) by H. Maturana and F.J. Varela (1979, 1985). The theory attempts to derive the entire phenomenology of autonomy from the property of all organisms to continually renew and in this sense auto-produce themselves. To stay with the car example, cars are made in factories, and if we continue the production line, cars will produce something else, namely, waste (which in turn will produce still other chemicals, etc.). So we have an open production diagram, factory-car-waste. In contrast, as we all know, our body was produced by another body, and the product (besides some waste) will be further bodies - and not just in the sense of parental propagation. It is calculated that every cell of the human body is replaced by another in about seven years. The cell dies, and a new one occupies its place. We have a closed production circle, body-body-body. Autopoiesis maintains that such a closed production circle has a new ability which is not present in any of its components. For example, if we have such systems in numbers and allow for them to change randomly, then, by the imperative of the existence of the descendants, the successions of these systems define their own range of interactions and their own range of perturbations. Only those interactions which are permitted by the autonomous design are possible. So, conclude the autopoiesis folks, the existence, and the evolution (if there is any) of these systems are consequences of their internal laws. Or, to reverse the argument, any kind of truly interesting autonomy and independence of supervision is only achievable by autopoietic systems, that is, by producing a machine that later continues to produce itself.

This is a tricky argument, which is far less trivially wrong than many people tend to believe. It is an inside-out Darwinian argument. Evolutionary theory puts the stress on how the environment acts on the organism by means of changing population composition. Here the idea is the opposite. Listen to this. As long as we have a system that has a basic organization (autopoiesis, mind you) that allows for it to produce offsprings that survive no matter what (for instance, by producing a variety of different offsprings which if survive the evolutionarist would call "adapted"), the structure of the system can be *anything*, and its properties beyond the said basic organization can be determined by various kinds of internal factors. To make it simpler: By introducing the trick of self-production (and variation), you buy yourself free to do your own work. It is like having a secretary who does all the dirty job (react to

environment etc.), and you can devote yourself to windsurfing, writing poems, chasing women (men), or determining input-output relations, if you wish.

Life is more complicated, however, and the autopoiesis idea was much criticized for its lack of coping with biological knowledge in the proper way. For instance, it fails to reflect on the adaptationism-exaptationism/constraints debate. Be it as it may, autopoiesis has at least one important suggestion: that autonomy cannot be a software concept, but one about the interplay between software and hardware. Software is always dependent. If we think about software in the production line metaphor, we see it immediately that no program can produce itself, nor can software alone make new hardware. (Programs can produce themselves only if someone buys more memory and ensures other elements of the environment.)

But this leads us away from our original topic of software agents. It is no accident, perhaps, that interest is increasing in systems that are self-sufficient in a real world sense, acting as incorporated agents situated in a rich physical environment, where independence from supervision has a more complete meaning and where evolutionary capabilities can be so important. The future is closer to science fiction than we thought.

## **Acknowledgement**

This paper is a written version of a tutorial lecture held on May 29, 1998 at CEU Budapest. The writing of the paper was supported by the research grants OTKA T25880 and MKM FKFP-0225. The supports are gratefully acknowledged.

## **References**

Ashby, W.R. 1964: Introduction to Cybernetics, Chapman and Hall, London.

Belgrave, M. 1998: The Software Agents Mailing List FAQ. Maintained by Marc Belgrave.  
[http://www.ee.mcgill.ca/~belmarc/agent\\_faq.html](http://www.ee.mcgill.ca/~belmarc/agent_faq.html)

Bergson, H.L. 1997: The Creative Mind : An Introduction to Metaphysics, Citadel Press, New York.

Bertalanffy, L. von 1976: General System Theory: Foundations, Development, Applications George Braziller, New York.

Brentano, F. 1973: Psychology from an Empirical Standpoint (International Library of Philosophy), Routledge, London.

Buckley, K.W. 1989: Mechanical Man: John Broadus Watson and the Beginnings of Behaviorism, Guilford Press, New York.

Chalmers, D. 1998: <http://ling.ucsc.edu/~chalmers/biblio.html>

Chisholm, R. M. 1986: Brentano and intrinsic value, Cambridge University Press, New York.

Dennett, D.C. 1989: *The Intentional Stance*, MIT Press, Cambridge, Mass.

Franklin, S. and Graesser, A. 1996: *Is it an Agent, or just a Program?: A Taxonomy for Autonomous Agents*. Proceedings of the Third International Workshop on Agent Theories, Architectures, and Languages, Springer-Verlag.,  
<http://www.msci.memphis.edu/~franklin/AgentProg.html>

Gardner, H. 1985: *The Mind's New Science*, Basic Books, New York.

Goodwin, R. 1993: *Formalizing properties of agents*. Technical Report CMU-CS-93-159, School of Computer Science, Carnegie-Mellon University, Pittsburgh, PA.

Heidegger, M. 1997: *Being and Time: A Translation of Sein and Zeit* (Suny Series in Contemporary Continental Philosophy and Culture), State Univ of New York Press, Stony Brook.

James, W. 1955: *The Principles of Psychology*, Dover, New York (original 1918).

Janca, P. 1995: *Pragmatic Application of Information Agents*. BIS Strategic Decisions, Norwell, United States.

Laing, R. 1977: *Automaton Models of Reproduction by Self-inspection*, *J.Theor.Biol.* **66**, 437-456.

Lindenfeld, D.F. 1980: *The Transformation of Positivism : Alexius Meinong and European Thought, 1880-1920*, Univ. of California Press, Berkeley.

MacKay, D. M. 1980: *Brains, Machines, and Persons*, Grand Rapids, Mich.

Maturana, H.R. and Varela, F.J. 1980: *Autopoiesis and Cognition*, D. Reidel, Dordrecht.

Merleu-Ponty, M. 1992: *Phenomenology of Perception*, Routledge, London.

Millikan, R.G. 1988: *Language, Thought, and Other Biological Categories : New Foundations for Realism*, Bradford Books, MIT Press, Cambridge, Mass.

Neumann, J. von 1966: *Theory of self-reproducing automata*. Edited and completed by Arthur W. Burks., University of Illinois Press, Urbana.

Putnam, H. 1960: *Minds and machines*, in (S. Hook, ed) *Dimensions of Mind*. New York University Press. Reprinted in *Mind, Language, and Reality* (Cambridge University Press, 1975).

Putnam, H. 1967: *The nature of mental states*, in: (Capitan & Merrill, eds.) *Art, Mind, and Religion*, Pittsburgh University Press. Reprinted in *Mind, Language, and Reality* (Cambridge University Press, 1975).

Quine, W.v.O. 1964: *Word and Object*, MIT Press, Cambridge, Mass.

Rashevsky, N. 1961: *Mathematical principles in biology and their applications*. Thomas, Springfield, Ill.

Rosen, R. 1985: *Anticipatory systems: philosophical, mathematical, and methodological foundations*, Pergamon Press, Oxford.

Sellars, W. 1979: *Naturalism and Ontology*, Ridgeview Pub., Reseda, Calif.

Turing, A.M. 1950: *Computing Machinery and Intelligence*, *Mind* 54, 236-245.

de Waal, F. 1982: *Chimpanzee Politics*, Harper and Row, New York.

de Waal, F. 1997: *Bonobo: The Forgotten Ape*, Univ. California Press, Berkeley.

Wooldridge, M. and Jennings, N.R.: *Intelligent Agents: Theory and Practice*, *The Knowledge Engineering Review*, Vol 10 (2), pp. 115-152, 1995.