# A Dialogue on Metasystem Transition

Valentin F. Turchin
The City College of New York

July 12, 1999

*This dialogue is between the author,* **T**, *and an imagined synthetic person,* **S**, *who has no definite positions of his own, but asks questions and makes judgements as, in the author's view, a typical reader could do.*

## 1 Epistemology

**T** Hello, **S**! A beautiful weather today, isn't it? Are you sure you want to discuss philosophy, instead of hiking and swimming?
**S** To tell the truth, I am not. I came here from curiosity, but I am not sure at all that we will not waste our time. Let us start. I reserve the right to say at any moment: that is enough, I am out.
**T** All right. I understand you very well. I also find many discussions on – and near – philosophy unproductive. Usually people simply do not understand each other; they speak different languages. There are many philosophical languages, and they change as time goes. As you know, I am not a professional philosopher, and to compare in detail the languages of various great philosophers is not within my competence. But I have always held the view that everyone must have his own philosophical language in which to answer the everlasting questions: What is the world? What am I? (the subject of *Ontology*). What is our knowledge of the world? How true is it? (*Epistemology*). What is Good and what is Evil? What are the supreme values and the meaning of life? (*Ethics*). Having such a personal language, one should be able to translate into it the ideas expressed in other languages.
**S** But, surely, it is not always possible to translate from one language to another. Remember the complementarity principle in physics. You can describe a quantum-mechanical particle in terms of its co-ordinate, or in terms

of momentum, but you cannot describe it in these two ways simultaneously. The more precisely you determine co-ordinate, the less you will know about momentum.

**T** I am afraid your example shows the opposite of what you intended to show. The incompatibility of the two descriptions holds only as long as you use the classical notions, and this was the point of Bohr's complementarity principle. To avoid problems, do not use classical notions where they are not applicable. In quantum mechanics the particle is described by its wave function. It can be written in the co-ordinate representation, or in the impulse representation, and it is easy to translate one into another by making the Fourier transform.

When we have two or more languages which partially describe a phenomenon, our goal should be to create a more complete theory which synthesizes and unifies the pre-existing theories. This is what has been achieved with quantum mechanics. I see no reasons why this should not be a typical case. I am against invoking the complementarity principle as a justification for the absence of a unifying theory. Maybe we simply did not work hard enough. I do not see any logical reason why a useful unifying language and theory cannot always be found. In the simplest case, different languages give us different projections of the same phenomenon, and can be easily combined, as when we have three projections of a moving particle on three orthogonal axes.

The reason why we do not always want to combine philosophical texts and languages is more down-to-earth: it is not that we cannot do it because of some universal complementarity principle, but that we simply do not need it. A text, or the whole set of texts written in a certain language, may express the meaning that adds nothing new, because we already know this and have expressed it in a different language. Or it may have no meaning at all. To make a discussion meaningful, we must make it sure that we understand each other. This is why I propose that we start our discussion with epistemology, to which the problem of meaning belongs. For some time I have been looking for a kind of a universal semantics, some guiding principle to understand a text in every possible language, if it, indeed, has any meaning.

**S** Did you find one?

**T** I think I did – to some extent. And I think every philosophy, and science as well, must start with the discussion of this, or a similar, principle. We need some criterion of meaningfulness. Otherwise we will not be able to distinguish between the meaningful and the meaningless. We simply will

not know what we are speaking about.

**S** I am eager to hear about the principle you discovered.

**T** Well, discover is too strong a word. My epistemology will not surprise anybody who is not unfamiliar with the modern philosophy. My semantic principle, briefly, is: the meaning of a linguistic object for me is in my ability to use this object as an instrument for making models of the world, in other words, in generating predictions about the world's processes. I come to this principle by arguing that whatever has meaning must, somehow, increase our knowledge, and the cybernetic idea of knowledge is that it is a model of reality.

Closely tied to this principle is the method in which I propose to develop philosophy: the method of *progressive formalization* [16] This is the method universally used in science. We first rely on an intuitive understanding of simple concepts, then on the basis of this understanding we convey the meaning of more formal and exact, but also more complex, concepts and ideas.

This statement itself is an illustration of my method. I used in it the words 'understanding', 'meaning', 'formal'. In due course, these notions should be analyzed and 'more formal and exact' meanings should be given to them, in their turn. These new meanings, however, will not come to *replace* the original meanings, but to *make an addition* to them.

Compare this with the situation in physics. We start this branch of science speaking about bodies and their masses, measuring distances in space by applying rulers, etc. Later, when we study the structure of matter, we find that those bodies and rulers, are nothing else but certain structures consisting of huge numbers of atoms. This concept of a ruler is, however, a new concept, even though it refers to the same thing. To come to the concept of a ruler as an atomic structure, we must pass a long path, at the beginning of which a ruler is a simple thing the usage of which is easy to explain.

In the Principia Cybernetica Project [5], we come to philosophy with the standards and methods of science. We try to define and explain such basic things as 'meaning', 'understanding', 'knowledge', 'truth', 'object', 'process' etc. But to explain, e.g., understanding, we must rely on understanding in its usual intuitive sense, because otherwise we will not know if we ourselves understand what we are saying; so, there will be little chance for our words to be meaningful.

Or take the concept of an object. In Principia Cybernetica we have a conceptual *node* devoted to it. But we cannot do without speaking about

objects long before we come to that node – in a close analogy with the two concepts of a ruler in physics.

Relations between things in this world are very often circular, so we are often at a loss when trying to start and finish definitions. Using various levels of formalization allows us to avoid vicious circles in definitions. Suppose we use informally some concept $A$ to define a concept $B$. Let us represent the fact that $A$ conceptually precedes $B$, or $B$ relies on $A$ as $A \prec B$. Then we want to make $A$ more exact: $A'$. We define it, and discover that it now depends on the already defined $B$. Hence if we were to require that in a formal definition of a concept all the concepts on which it relies are formally defined, we would either have to limit ourselves to strictly hierarchical subsets of concepts, or never finish the job, moving in a vicious circle. Instead, we recognize that there are various levels of formalization of essentially the same concept, and we allow them to coexist. Thus after defining $B$ with the use of $A$, we define $A'$ using the informal concept $B$; since $B$ relies on $A$, the old, informal version of $A$ is not discarded, but stays in the system of concepts. Now we could make the definition of $B$ more formal, basing it on $A'$ instead of $A$; on the next turn of this spiral, we may wish to define even more formal concept $A''$, etc.:

$$A \prec B \prec A' \prec B' \prec A'' \prec B'' \ldots etc.$$

Whenever we want to understand a definition, we start unwinding the chain of dependent definitions from right to left, until we come to basic intuitive notions about which there should be no disagreements.

**S** You define your primitive concepts using the concept of modelling. But this concept itself is far from primitive. It relies on the same primitive concepts which your are defining.

**T** Yes. This is the process of progressive formalization. I define modelling by appealing to your understanding of the basic method of science. After that I start defining various concepts of philosophy referring to something you already understand: modelling. I define the place of these concepts in modelling. It gives you a way to decide if in a given context these concepts are used properly. This means that my definitions are more formal than if the concepts were not defined, but simply described and announced primitive.

**S** But the concept of modeling is quite advanced. Why should you take it as the beginning? I may not believe in the model epistemology, but agree with your ontology that actions are primary reality, and accept the idea of progressive formalization. Starting from such primitives, I would come to

formal definition of modelling. But you insist on accepting epistemology first. This only makes things more difficult for me.

**T** You are free to start from any point in the spiral of progressive formalization. But if this point is not what I take for the beginning, you have to rely on the intuitive understanding of abstract philosophical concepts. With different people it may be different; only the words used are the same. I do not know how to compare intuitive meanings. But I know how to check that a person uses the idea of modelling correctly. Therefore, the explanation of abstract concepts in terms of the concept of modelling becomes, for me, acceptable. This is why I start with epistemology. But I repeat that you can start the discourse from any point. If you wish, start it with ontological primitives. I start with epistemological primitives.

**S** I expect that you will now explain, or, as you say, make more formal, what is a model, and what does it mean that it is formal or informal.

**T** Exactly. First, about modeling. It is a kind of activity of a cybernetic system, in particular, a human being.

**S** And what is a cybernetic system?

**T** No comment. I believe that whatever notion of a cybernetic system you have, it will do. In due time, in this dialogue, or elsewhere in the course of the Principia Cybernetica Project, we shall give an answer to this question. But not now. This is the method of progressive, gradual, formalization.

**S** A convenient method, indeed! I could go on insisting that you give a definition now.

**T** And block any further discussion. This is easy to achieve by various means. To your irony I answer: yes, it is convenient. It allows to have things started.

We can construct models of various systems. Let me call the system we are modeling simply 'the world', meaning by that some part or aspect of the world as we see it. The system that constructs the models, to which I have been referring until now as 'we' or 'I', will be, in the third person, called *the subject of knowledge*. The model we discuss is a subsystem of the subject of knowledge.

The most immediate kind of a model is a system that implements the concept known in mathematics as *homomorphism*. This system can be described as follows (see Fig.1)

Let $W_1$ be a state of the world as reflected in the primary sense organs of the subject of knowledge. Let $R_1$ be the *representation* of the state $W_1$. By this I mean the existence of some procedure $M$ (mapping) which produces $R_1$ when $W_1$ is given: $M(W_1) = R_1$. Suppose further that the subject

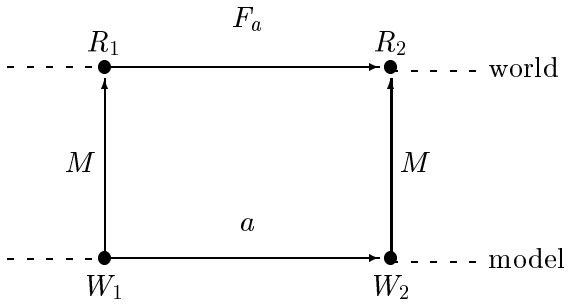


Figure 1: The scheme of modeling

of knowledge takes an action $a$. As a result, the state $W_1$ changes into $W_2$. (Among possible actions of a cybernetic system there is the action of doing nothing: just waiting for a period of time). To be a model, the system must be able to perform one more procedure, let us call it $F_a$. It mimics in the model the effect of the system's action $a$ in the world, so that $F_a(R_1) = M(W_2)$. Thus by applying $F_a$ to $R_1$ the system can predict, to some extent, the development of events if it takes action $a$. Then it can choose an action which helps it survive. Modeling is a powerful instrument of survival, and this is how it emerged in the course of evolution.

**S** I must note that your concept of a model is not the only one. For example, if your mapping procedure, which implements a *function*, is replaced by a general *relation*, that will be again qualifying as a model, and you will find such a definition in some books.

**T** Yes. But I have serious reasons to choose my definition. I will discuss this later, when we come to the evolutionary origins of knowledge.

The concept of modeling as I have defined it can be generalized by declaring a model any tool which produces predictions. My definition of a prediction is: a statement that a certain process is finite, meaning by being finite that it comes to a certain, specified in advance, stage. In particular, the prediction supplied by the above-described model, namely $F_a(M(W_1)) = M(W_2)$ is nothing else but the 'finiteness' of the process which we shall denote as $P$ and which can be described as follows. Apply $M$ to $W_1$, then apply $F_a$ to the result, and call $X_1$ the result of that. Let the cybernetic system that carries the model make action $a$. Let the resulting state of the world be $W_2$. Apply $M$ to $W_2$ with the result $X_2$. Apply the

comparison process to $X_1$ and $X_2$. We define comparison as a process which stops when (and if) the identity (or equivalence) of $X_1$ and $X_2$ is established. Thus a successful end of this process means a successful end of the whole process $P$. Therefore, the statement $F_a(M(W_1)) = M(W_2)$ is a prediction that $P$ is finite.

Predictions are, in principle, verifiable. You only have to initiate the process that it is about and wait until it comes to the final state.

As you remember, I started by tying up meaning to the cybernetic concept of knowledge. A model, or a generator of predictions, does certainly represent knowledge. However, we must not limit the whole concept of knowledge to a generator of predictions. Pieces of our knowledge (propositions) do not necessarily produce verifiable predictions, but may produce something which will produce predictions. Moreover, they may produce objects which produce objects which produce predictions, and so forth to any height of the hierarchy of knowledge objects. I will often refer to this process as *hierarchical production* of predictions. A simple example from mathematics: the equation $x + y = y + x$ is not immediately verifiable, but it produces such an equation as $7+4 = 4+7$. This statement, in its turn, is still too abstract for a direct verification. We can, however, verify the prediction that four apples and seven apples can be added in either order with the same result. If we take something even more abstract, like Maxwell's equations, we shall see even a longer hierarchy of specification before we come to observable facts.

I propose, therefore, the definition: a piece of knowledge is an object which we can use for hierarchical production (or generation) of predictions [16]. In a more formal way: a piece of knowledge is a generator of predictions or other pieces of knowledge. This recursive definition allows a piece of knowledge to produce a hierarchy of objects before it starts producing predictions. Note that according to my definition a thing may never start producing predictions, and still qualify as knowledge: call it empty knowledge. The reason for the inclusion of this case is that with a recursive definition of generating procedures we cannot always tell in advance if a given generator will produce a single object.

Now I have come to the point where a more formal definition of *formal* is due. A statement or a language is formal if its usage relies only on the 'form' of linguistic objects, and not their intuitive meanings.

**S** But *whose* usage it is?

**T** A good question. My next step in making this definition more formal and precise is to specify a set of perceptions and actions which are regis-

tered and performed in the same way by all members of the society whom the languages serves. Let us refer to these perceptions and actions as *universally defined*. A language is formal if the processes involved in its usage, namely the representation function $R(w)$ and the modeling function $M(r)$, are expressed in terms of universally defined perceptions and actions. The notion of universally defined, though, cannot be formally defined. Thus, the difference between formal and informal always remains informal.

We usually assume that universally defined perceptions and actions can be relegated to a machine. The question is still open whether this is a realistic assumption. We accept it with a qualification that if there is a doubt about a specific abstraction or action, it must be excluded from the universally defined set. Then a formal language is a language usable by a properly constructed machine. A machine of that kind becomes an objective model of reality, independent of the human brain which created it. Science is construction of such machines.

**S** I understand, this is the reason for your program of progressive formalization.

**T** Exactly. We create formal versions of our common notions in order to understand better how our language and mind work, and to create artificial languages and minds, which will imitate our mental processes, and one day, perhaps, go beyond what is possible for us. By a series of consecutive formalizations, philosophy becomes science.

Thus let us continue on this path. Our definition of knowledge allows me to further define what is meaning and what is truth. When we state something we, presumably, express our knowledge, even though it may be hypothetical or false. Thus to be meaningful, a proposition must conform to the same requirement as a piece of knowledge: we must know how to be able to produce predictions from it, or produce tools which will produce predictions, or produce tools to produce such tools, etc. If we can characterize the path from the statement to predictions in exact terms, the meaning of the statement is exact. If we visualize this path only vaguely, the meaning is vague. If we can see no path from a statement to predictions, this statement is meaningless.

**S** You cannot say just "meaningless". It may be meaningless *for us*, but will it forever remain meaningless for everybody?

**T** True enough. This is why I said "if *we* can see no path". What I want to emphasize is not the subjective side of all knowledge (about which there is a general consensus nowadays), but the specific mechanics of acquiring a meaning: production of verifiable predictions. A piece of knowledge is true if

the predictions made by the user of knowledge on the basis of this knowledge come true. Since there is no general method to determine if a recursive generator produces a result of a given kind, there is no generally applicable method to establish truths. Since sets of predictions, like multidimensional vectors, are hard to compare, there is no universal evaluation of truths.

Remember, you asked if I have found a universal semantic principle to decide on meanings, and I said 'to some extent'. My reason for being cautious is that we usually expect from such a principle that it guarantees a definite answer with respect to any question. As I have just said, there is no such principle to decide on truth, even if the statement is formal. As for the meaning, the universal principle exists if we limit ourselves to formal languages. It requires that by our construction of the statement, it is a machine which produces only predictions. However, when we push forward the frontier of theoretical knowledge, we deal with informal statements which cause flows of ideas in our heads but are not (yet!) ready for formalization as machines. There is no formal principle to judge on the validity of such statements other than wait until they yield predictions. My semantic principle only indicates the goal, but cannot offer a universal algorithm.

But I believe that this semantic principle, nevertheless, can improve mutual understandability in philosophical and methodological arguments, because it indicates the direction in which to look for resolution of conflicts: it is how what we say translates or may translate into production of predictions. I am trying to show this in our present discussion. I formulate whatever I have to say either as a model of reality, or as a way leading to construction of models. Thus I see my own philosophy as definitely meaningful.

I propose this as a general guiding principle in human attempts to understand each other. If the other side in a dialogue produces chains of words the meaning of which you cannot grasp, ask it to explain how these words are relevant for construction of the world's models. I believe, optimistically, that if both sides hold to this method, the discussion will become more meaningful.

**S** What is the meaning, in your theory, of the statement: The distance from Boston to Portland is 107 miles. Is it formal or informal?

**T** It is the prediction that the following process comes to a successful end: set the odometer in your car at zero, drive from Boston to Portland, and compare the figure at the odometer with 107. I believe this instruction is completely within the universally defined perceptions and actions. So it's meaning is formal.

**S** But my statement is more abstract. It does not include a specific indication at the procedure of measuring. I could go from Boston to Portland by foot.

**T** If you associate the concept of distance with more than one method of measurement, the statement of the equivalence of various methods is implicit.

**S** Do you seriously believe that in this way you can interpret the meaning of *any* statement which we can express in a natural language? Even, say, from *Nursery Rhymes*?

**T** Yes.

**S** OK. 'Mary had a little lamb'.

**T** Well, sentences of natural human languages are burdened with many different implication, often conflicting. But I can sketch how your sentence can be interpreted in terms of prediction-making.

First of all, we deal here with the past tense. Which means that our statement does not directly produce predictions, but adds to what can be called an internal picture of the world, which every person has.

**S** Are not you retreating from your original position that all that has meaning is prediction generation?

**T** Not in the least. The personal picture of the world is part of prediction generation, and has meaning to the extent it helps predict. Remember the modeling scheme? We viewed $F_a(R_1)$ as a function of the current state of the world $R_1$ and the parameter $a$, the action the subject system could take. But this function depends also on our mental picture of the world as one more parameter.

**S** Then what you call the picture of the world is nothing but memory.

**T** Almost. It is that part of memory which is relevant for prediction making. Every experience adds something to your memory, but this addition may or may not be meaningful. If I say to you: 'Aderti was compy stallous yesterday', the fact that I said this may stick in your memory, but the sentence itself will add nothing to your ability to make predictions. So I say that the sentence is meaningless.

Because of the human ability to have and *construct* mental pictures of the world – the faculty of imagination, to which we shall return once again – we can treat mental pictures the same way as we treat reality. In particular, we can make 'predictions' about events in our pictures which, of course, will not be predictions proper but some constructions in those pictures. If Julius Caesar in our mental picture drops an apple, we assume that it falls down, and this becomes one more element in our picture of the world. If we know

that Mary had a little lamb, we can assume that she gave it some food, and it was not hamburgers and beer. In this way we reduce the meaning of past-tense texts to the meaning of texts about the present.

So, 'Mary *has* a little lamb'. Now we face a problem that is known in computer science as *knowledge representation*. The standard method is to decompose a natural language statement into a formula of the predicate calculus using some *primitive* predicates. In our case this translation may be:

$$\exists x, y[\text{Person}(x) \wedge \text{Called-Mary}(x) \wedge \text{Little-lamb}(x) \wedge \text{Has}(x, y)]$$

To translate back into English: 'There exist such objects $x$ and $y$ that $x$ is a person called Mary, $y$ is a little lamb, and $x$ has $y$.

Primitive predicates are defined by appealing directly to our human perception, and the predicate is true if and only if our perception – which is a certain process of verification – comes to a successful end. For example, in order to establish that Little-lamb(x) is true, i.e. some object $x$ is a little lamb, and not a big bad wolf, we just observe $x$ and confirm that we see a lamb. The meaning of the statement Little-lamb($x$) is in the prediction that the verification process ends successfully. Existential quantification, i.e. the statement 'there exists such $x$ that ... etc.' is also understood as a prediction, namely the prediction that if you start examining all the objects in the Universe – in fact, on the Earth (this is clearly assumed in the sentence) in search of an object $x$ which is a person and meets all other requirements, then you will sooner or later find it (and stop). The prediction is that this search is finite.

Sentences of natural languages will never allow to define their meaning in a completely formal way; not until we decompose human thought and soul into billions of billions of elementary units, which may or may not be possible, we do not know yet. But we can move (almost infinitely) in the direction of greater precision and formality. We can write a program which will distinguish between an image of a lamb and that of a wolf. To check that $x$ is called Mary, we can refer to her birth certificate, or observe that $x$ answers when addressed as Mary, etc. As in the case of distance measurement, the full definition of a concept should include all relevant tests, and a mechanism to decide on the answer when there are disagreements between ... I see that you look wistfully through the window.

**S** Yes, it may be a good idea to have a swim.

**T** Very well. I only want to make a few remarks to finish with epistemology.

First, my theory of meaning leads to a theory of the value, or usefulness, of information. Shannon's measure of information does not include this aspect. Obviously, one can receive huge amounts of information measured in bits and make no use of it at all. We often hear the question: how to measure *useful* information? My answer is: in the last analysis, information in any message is meaningful, or useful, to the extent it is used for making predictions. Basically, this is the same concept of knowledge and meaning that we have been discussing today. Information is useful to a cybernetic system if it enhances its knowledge, otherwise it can be thrown out as trash.
**S** You reduce meaning to knowledge only, which in your theory is generation of predictions. But what about passing useful instructions? What about skills, the know how? You cannot deny that such instructions have meaning. But I cannot see generation of predictions there.
**T** Note that you used the words *know how*, thus treating skills as knowledge, which is quite correct. In my definition of a successful, finite process, you can distinguish two parts: the process proper – let us denote it as $P$, and the test $T$ which determines if the stage reached is final, i.e. satisfying the pre-set requirement. When we want to find a set of predictions, one of the following two cases usually holds. First, we can specify $P$ and then ask what will happen, i.e. which kind of tests $T$ will be finally successful when following $P$. This is the most direct meaning of the word 'prediction'. But equally important is the second case when we specify $T$ and ask what kind of process $P$ will lead to the desirable result. This is your case of useful instructions. The essential content in both cases is the same: that the process $PT$ is finite.

My second remark is that I have tested elsewhere the validity of my approach to knowledge, meaning and truth by applying it in the field which does not allow imprecision and vagueness but requires a complete formalization and unambiguity, – in mathematics. I have done this in [14], where a 'cybernetic' foundation of mathematics is developed, based exactly on the principle that the meaning of mathematical statements is only in recursive generation of predictions expressed in a formal language. This approach gives answers to the classical questions about mathematics; in particular, it gives a new and constructive interpretation of set theory.

Finally, I have given you no more that a very brief introduction to the way I propose to treat the problems of epistemology. Many aspects of these problems I have left out, for example, the treatment of possibility, as in 'it may be that ...'. Many other aspects are not yet elaborated at all, and I hope to have a chance to work at those in the frame of the Principia Cybernetica

Project.

## 2 Metasystem Transition

**S** It is time to start discussing the concept of metasystem transition, which is, after all, the goal of our meeting.
**T** Yes, but in a moment I will have to make one more journey into philosophy.

In *The Phenomenon of Science* ([12]) I define metasystem transition as follows (see Fig.2). Imagine a system $S$ of some kind. Suppose there is a way to make a number of copies from it, possibly with variations. Suppose that these systems are united into a new system $S'$ which has the systems of the $S$ type as its subsystems, and includes also an additional mechanism which controls the behavior and production of the S-subsystems. Then we call $S'$ a metasystem with respect to $S$, and the creation of $S'$ from $S$ a metasystem transition (*MST* for short).

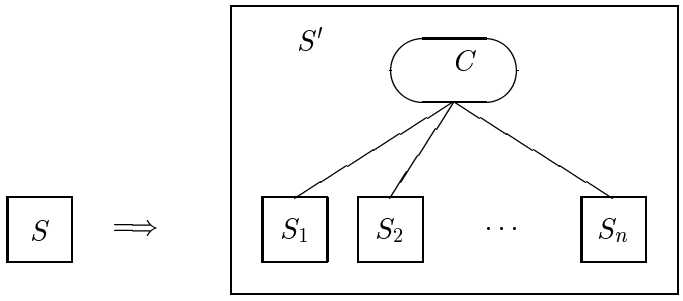


Figure 2: Metasystem transition

As a result of consecutive metasystem transitions a multilevel structure of control arises, which allows complicated forms of behavior. I show, further, that the major steps in evolution, both biological, and cultural, are nothing else but metasystem transitions of a large scope. The concept of metasystem transition allows us to introduce a kind of objective quantitative measure of organization and distinguish between evolution in the positive direction, progress, and what we consider an evolution in the negative direction, regress. In particular, I offer an interpretation of one of the most important aspect of the biological evolution: the appearance of human thinking and human society.

**S** I am distrustful about the notion of *progress* because it is value laden – and in a very skewed manner at that – in our Western culture. I merely observe a *progression* toward complexity.

**T** This is your right, of course. But in my value system this progression and the emergence of man, in particular, come with the sign plus. So I call it progress.

*The Phenomenon of Science* was written some twenty years ago. Since then I had a chance to read more literature on cybernetics and evolution, and I discussed the concept of metasystem transition with various people in various contexts. I am convinced more than ever that mine is a valid way of seeing the evolution of the world and predicting its future. But I feel a kind of necessity to make the concept of *control* more definite and precise. In cybernetic literature this concept is often identified with a very specific scheme, which I prefer to call a *regulation* scheme, where the metasystem's purpose is to keep a certain variable constant; see, e.g. [10]. When I speak of a hierarchy of control, I understand control in a very general sense, which includes the classical regulation scheme and any ways of duplication, variation, integration, manipulation, exploration etc. For example, the creation of the language of formal logic to make mathematical proof into a mathematical object is a typical MST, although it cannot be reduced to a regulation scheme.

So it seems to me that there must be a way of defining and using the MST concept with a concept of control which is very general and fundamental, one of the main features of being; then evolution by metasystem transitions will also become an inalienable feature of the world. But to define such a concept we need the context of *ontology*, the part of philosophy which is called to tell us what does it mean *to be*, and what, in the last analysis is the world. To define control, I want first to define being.

**S** Is it really necessary?

**T** We could discuss the world's future without it. But the work I am doing is part of the Principia Cybernetica Project, and its purpose is to create an all-embracing, complete philosophical system on the basis of cybernetic ideas. We want to make this basis into a system of conceptual nodes which could be then used both for construction of intelligent machines, and, possibly, creation of new scientific theories. Also, if we demonstrate that our concepts form a consistent and complete picture of all that is, it will make our conclusions about future more convincing. As you remember, I insisted that we start with epistemology and the principle of progressive formalization. We discussed why we needed progressive formalization. Now we shall

discuss how to start it.

**S** Wait a minute. You jumped from ontology, which is to me more or less the same as metaphysics, to formalization and intelligent computers. I still do not see the need for you to drown in the bog of metaphysics.

**T** I said: to start progressive formalization. Metaphysics is often viewed as something opposite to physics and utterly useless for any reasonable purpose. This attitude is a hangover from outdated forms of empiricism and positivism, namely the naive reflection-correspondence theory of language and truth, which sees language as an image, a replica of the world. It is easy to conclude from this theory that any expression of our language which cannot be immediately interpreted in terms of observable facts, is meaningless and misleading. This viewpoint in its extreme form, according to which all unobservables must be banned from science, was developed by the early nineteenth-century positivism (August Comte). From this perspective, metaphysics is definitely meaningless.

But our view of language and truth is different. We understand language as a hierarchical model of reality, i.e. a device which produces predictions, and not as an image of the world. This device, especially in its higher levels of structure, need not 'look like' the things it is about; it only should produce correct predictions. Therefore, the claim made by metaphysics is now read differently. To say that the real nature of the world is such and such means to propose the construction of a model of the world along such and such lines. Metaphysics creates a mental structure to serve as a basis for further refinements. Metaphysics is the beginning of physics; it provides foetuses for future theories. It may take quite a time to translate metaphysics into an exact theory with verifiable predictions. Before this is done, metaphysics is, like any fetus, highly vulnerable. But we need *some* metaphysics. On our agenda is the creation of universal models of the world, which would allow us, in particular, to interpret human thought expressed in natural language. How should we start this enterprise? What concepts must be taken as the basis? This is the same as to ask: what *is* the world? What is its ultimate essence? It is the business of metaphysics to give answers to these questions.

**S** So, what is the ultimate essence of the world?

**T** My answer is: *action* [17]. Which means that it is action that must be taken as the ultimate building element in the construction of world models. This is a truly cybernetic approach. Physics is concerned with the material of the world, the matter-energy aspect of it. Cybernetics abstracts from the material and concentrates on control, communication, information. All of these are actions.

Intuitively, we see the world as a collection of objects occupying some space and changing in time. Objects are seen as primary, change as something secondary, which could or could not take place. I reverse this relationship. I modify the famous Schopenhauer's formula as

*The world is action plus representation*

with action taking ontological precedence over representation.

**S** For Schopenhauer it was *will*, not action.

**T** Yes. But the two concepts are rather close. If I understand Schopenhauer correctly, will is a universal factor that makes action possible. Will manifests itself as action. Taking action as the basis, I get closer to our usual perception of the world, yet far enough not to treat physical objects as the 'true' elements of reality. Objects are representations of the world in our mind. They come into being through sensations. But sensations do not exist as objects; they are *actions*, a form of *interaction* between the subject of knowledge and the rest of the world.

**S** I do not understand your ontological precedence of action over anything else. I would rather understand Schopenhauer's will as existent. At least, will is something definite, permanent. The quality of permanence is necessary for being in existence. That action can exist seems to me a contradiction, a logical absurdity.

**T** Here we face the most intriguing part of metaphysics: the concept of 'real existence'. Our cybernetic epistemology, according to which all meaningful statements are hierarchical models of reality, has a double effect on the concept of existence. On the one hand, theoretical concepts, such as mechanical forces, electromagnetic and other fields and wave functions, acquire the same existential status as the material things we see around us. On the other hand, quite simple and trustworthy concepts like a heavy mass moving along a trajectory, and even the material things themselves, the egg we eat at breakfast, become as uncertain and open to discussion as theoretical concepts.

One could argue that there is simply no need in the concept of real, or ultimate, existence, because all theories, in the last analysis, explain and organizes observable facts, which all are, and will always be, facts of our perception. This is formally true. But we still do feel a need for our theory to start with such basic entities that their existence is impossible to deny. Somehow it seems that such a theory has better chances for success.

You require permanence for things that exist. But you know that there is nothing really permanent in this world. It seems to you that there is a

logical contradiction between action and existence because from the beginning, subconsciously, you identify existence with being an object. When I define existence as a feature of a theory of the world, this contradiction disappears. Thus I take the concept of action *in abstracto*, and on this basis try to interpret the fundamental concepts of our knowledge: what are objects, what is objective description of the world, what is space and time, etc.

**S** You did not yet define what is representation.

**T** Sure. You remember that according to our epistemology every meaningful statement is a model of reality, a dynamic entity. There are certain correspondences between the actions of the model and the actions in the real world: the former *mimic* the latter. All the rest in the statement, is representation. A statement is made significant by the actions involved in it; the representations used are secondary. Two models may be similar but based on completely different representations, as when we compare analogue and digital computation. While actions in our models reflect actions elsewhere in the world, our representations reflect nothing; they have no meaning of their own.

**S** So, representations are objects? Passive?

**T** Usually we see them as objects. But the concept of an object itself is not independent of actions; it is only an expression of a certain stability in relations between actions.

**S** Mmm...

**T** I see that I must explain what in my metaphysics is an *object*. Suppose I am aware of a tea-pot on the table in front of me. I recognize the image on my retina as belonging to a certain set of images, the abstraction 'tea-pot'. But there is more to it. I perceive the tea-pot as an *object*. The object 'tea-pot' is certainly not a definite image on the retina of my eyes; not even a definite part of it. For when I turn my head, or walk around the table, this image changes all the time, but I still perceive the tea-pot as the same object. The tea-pot as an object must, rather, be associated with the transformation of the image on my retina which results from the changing position of my eyes. This is, of course, a purely visual concept. We can add to it a transformation which produces my tactile sensations given the position and movements of my fingers.

The general definition of an object suggested by this example consists of three parts.

(1) First we define a set $R_{ob}$ of representations which are said to represent the same object; in our example this set consists of all images of the tea-pot

when I look at it from different view-points, and possibly, my sensations of touching and holding it.
(2) Then from the set of all possible actions we separate a subset $A_{cogn}$ of actions which will be referred to as *cognitive*; in our case $A_{cogn}$ includes such actions as looking at the tea-pot, turning my head, going around the table, touching the tea-pot etc. – all those actions which are associated with the registration of the fact that a tea-pot is there.
(3) Finally, we define a family of functions $f_a(r)$, where for every cognitive action $a \in A_{cogn}$, the function

$$f_a : R_{ob} \to R_{ob}$$

transforms a representation $r \in R_{ob}$ into $f_a(r) = r'$ which is expected as a result of action $a$.

The most important part here is the third; the first two can be subsumed by it. We define an object $b$ as a family of functions $f_a$:

$$b = \{f_a : a \in A_{cogn}\}$$

The set $A_{cogn}$ is the domain of the index $a$; the set $R_{ob}$ is the domain and co-domain of the functions of the family.

When I perceive an object $b$, I have a representation $r$ which belongs to the set $R_{ob}$; I then execute some cognitive actions, and for each such action $a$ I run my mental model, i.e. perform the transformation $f_a$ on $r$. If this anticipated representation $f_a(r)$ matches the actual representation $r'$ after the action $a$:

$$f_a(r) = r'$$

then my perception of the object $b$ is confirmed; otherwise I may not be sure about what is going on. Observing a tea-pot I check my actual experience against what I anticipate as the result of the movements of my head and eyeballs. If the two match, I perceive the tea-pot as an object. If I travel in a desert and see on the horizon castles and minarets which disappear or turn topsy-turvy as I get closer, I say that this is a mirage, an illusion, and not a real object.

The concept of an object naturally (one is tempted to say, inevitably) arises in the process of evolution. It is simply the first stage in the construction of the world's models. Indeed, since the sense organs of cybernetic animals are constantly moving in the environment, these actions are the

first to be modelled. In the huge flow of sensations a line must be drawn between what is the result of the animal's own movements, and the other changes which do not depend on the movements, are *objective*. Looking for objectivity is nothing else but *factoring out certain cognitive actions*. Function $f_a$ factors out the action $a$ by predicting what should be observed when the only change in the world is the subject's taking action $a$. If the prediction comes true, we interpret this as the same kind of stability as when nothing changes at all. The concept of object fixates a certain invariance, or stability, in the perception of a cybernetic system that actively explores its environment.

**S** Still I find it difficult to accept your view. It goes against the whole of modern science, according to which the world exists as a collection of objects, while actions are transitions between states of the world.

**T** But I do not reject this approach, I am perfectly ready to go along. The question is: what are those states? You consider them as something primary. I go further and define a state of the world as the set of all actions that can take place in this state. If these sets are identical then the states are identical. Note that in this way I reduce two basic concepts, action and state, to one: action. You cannot do the same taking leaving only state. Action as a change of state is a new concept; change cannot be expressed in static terms, as we have known starting with Zeno's paradoxes. So, on the purely logical reasons we are tempted to accept action as the only foundation of the world.

Now, consider this in the context of physics. According to our present understanding of the world, all the variety of events we observe result from elementary acts interactions between elementary particles. These acts constitute unquestionable reality, while both our theory, and our intuitive picture of the world, are only representations of reality. Furthermore, it is the physical quantity of *action* that is quantized by Plank's constant $h$. This can be seen as an indication that action should have a higher existential status than space, time, or matter.

**S** Well, it is not immediately clear whether the concept of action as we understand it intuitively and the physical quantity that has the dimension of energy by time and is called 'action' are one and the same, or related at all.

**T** This is true. That the physicists use the word 'action' to denote this quantity could be a misleading coincidence. Yet the intuitive notion of an action as proportional to intensity (intuitive understanding of energy) and time does not seem unreasonable. Furthermore, it is operators, i.e., *actions*

in the space of states, that represent observable (real!) physical quantities in quantum mechanics, and not the space-time states themselves!

Even if we reject these parallels and intuition as unsafe, it still remains true that neither space, time, nor matter are characterized by a single constant omnipresent quantum, but a combination of these. Is it not natural to take this combination as a basis for the picture of the world — if not for a unifying physical theory?

**S** It may be.

**T** What concepts have we already defined in our metaphysics of action?

**S** Representation, object, and state.

**T** Good. Now I want to define agent, freedom, and related concepts.

When we speak of an action, we speak also of an *agent* that performs the action. Formally, we can define an agent as a set of actions which is organized both sequentially and in parallel. We say then that every action from this set is performed by *the same* agent. In a given state of the world there may be many possible actions for a given agent. We say that this agent has the *freedom* to choose between them. When one agent's action restricts the freedom of another agent, we speak of *causation*. In the extreme case no freedom may be left to the agent; such an agent is referred to as a (deterministic) *machine*.

**S** Why do agents only restrict other agents? Why you exclude the cases where an action *increases* the freedom? For example, you can let somebody out of jail, thus increasing his freedom.

**T** Note, however, that I let the guy out by restricting the freedom of locks and jailers to keep him inside. I think this is a general rule. Whenever an action increases freedom, it does so by restricting restrictions.

Agents are, of course, representations, not actions. But we distinguish them from passive objects. We break down all representations into agents and objects. Both agents and objects are defined, in the last analysis, by actions; agents – by those actions they perform, objects by the actions through which they are perceived, as those I denoted $f_a(r)$ above.

**S** You defined agents as sets of actions. Now you say that agents are representation, not actions. Is this not a contradiction?

**T** . No. From a purely formal set-theoretical point of view, a set of actions is not an action itself. I used the set-theoretical language in order to give a concise definition, as they do in mathematics. You remember that one of the definitions of a real number in calculus is that it *is* a set of rational numbers. Here *is* does not mean *is the same*, as you may see the same girl on two different pictures. A real number is a concept on its own, an element of

our prediction machines. A set of rational numbers defines one real number as different from others.

Now I introduce an important relation between agents and objects, which I will call, following [8] and [6], the *semantic relation*. The action which an agent $A$ is about to perform very often depends on a certain object $b$. We shall call $b$ a *code*, agent $A$ its *interpreter*, the action of $A$ *interpretation*, and the relation between $A$ and $b$ a semantic relation; we say that $b$ *informs* $A$. Often we want to distinguish between the object $b$ and the *information* it carries. Information is an abstraction from the object in a semantic relationship, where only those features are left which have bearing on the actions of $A$. Thus two texts carry the same information for the reader if they different only in the font they are set in.

I believe that the existence of semantic relations is such a fundamental feature of the world that it cannot be reduced to, or defined through, anything more primitive. If there were no semantic relations, there could be no objects in our experience, because the perception of an object is its interpretation.

**S** Aha! You return to the objects their reality through semantic relations.

**T** I never doubted the reality of objects. But they are secondary to the primary reality of actions. An object is a code. An interpreter is an agent. These are two aspects of the action. Both are representations, and to some extent, are arbitrary. We often can alter the code without changing the action of the interpreter. And we can define the same action using a different code-interpretation pair.

As we discussed, modeling is a dynamic process. Introducing agents we make the first real step towards construction of the world's models. Our self-consciousness plays a decisive role in this step. Among all agents there is a special one: the agent denoted by the first person pronoun 'I'. This is the only agent of which we know from within: by *performing*, not *perceiving* actions. When we speak of our actions, or actions of other human beings, we know very well what the agent is: just the person whose action it is. We reconstruct this notion, of course, by extension from our own 'I'. When we speak of such animals as dogs, we again have no doubt in the validity of the concept *agent*. This reasoning can be continued down to frogs, worms, amoebas, trees, and inanimate objects, without any convincing arguments for stopping. When we say: 'the bomb exploded and the ship sank', are there any reasons to object against understanding this in the same way as if we were speaking about people and dogs? After all, the bomb might not explode, and with a given explosion the ship might or might not sink, de-

pending on the *ship* itself, the ship as a whole. Notice that even given a definite bomb and a definite ship, the result might not be uniquely predetermined.

And what about an *act* (sic!) of radioactive decay? It is definitely an action, but whose action is it? The physicist could say that the agents here are electrodynamic and chromodynamic fields. This makes sense because of the theory the physicist has. If we do not have such a theory, we simply say that there is a special agent for each possible act of radioactive decay. At each moment in time this agent makes a choice: to decay or not to decay. This immediately explains the exponential law of radioactivity.

**S** This is an anthropomorphism, which has been obsolete for hundreds of years.

**T** Since the primary instance of an agent for a human being is oneself, it is not surprising that in primitive societies the concept of agent is understood anthropomorphically: a 'spirit' which is very similar, if not identical, to ourselves.

The development of modern science banned spirits from the picture of the world. But agents, cleared from anthropomorphism, still remain, even though the physicists do not call them so. What is Newtonian force if not an agent that changes, every moment, the momentum of a body? Physics leaves – at least at present – the concept of agent implicit. We need it explicitly because our metaphysics is based on the concept of action, not to mention the simple fact that cybernetics describes, among other things, the behavior of human agents.

Speaking of a human being, we call the topmost agent its *will*. The freedom of this agent is the freedom of will.

The concept of will assumes the existence of *freedom* to exercise the will. Thus recognizing will as a cornerstone of being, we do the same for freedom. For the mechanistic world view of the nineteenth century freedom of will was a misconception, a nuisance which escaped satisfactory definition within the scientific context. For us the human freedom of will is a necessary element of the world order.

There is genuine freedom in the world. When we observe it from the outside, it takes the form of quantum-mechanical unpredictability; when we observe it from within, we call it our free will. This freedom is the very essence of our personalities, the treasure of our lives. Logically, the concept of free will is primary, impossible to derive or to explain from anything else. The concept of necessity, including the concept of a natural law, is a derivative: we call necessary, or predetermined, those things which cannot

be changed at will.

Now I can define finally define process, system, control, and metasystem transition.

A *process* is simply a collection of actions which are joint sequentially or in parallel: a complex action.

A *system* is a collection of agents and objects, thus it is a complex representation. This is a pretty general definition. It checks against various definitions from the systems theory literature, though. For example, one would often define a system as a collection of objects and relations between them. In our terms, the objects will be passive representations, while the relations will be seen as agents which initiate decision processes which determine whether a relation is held (I believe only in constructive definitions).

The concept of control is presented in Fig.3, where system $S$, composed of some agent(s) $A$ and representation(s) $R$, is controlled by the system $C'$.
**S** But agents are at the same time representation, so I do not understand the breaking of $S$ into $A$ and $R$. Maybe you wanted to say *objects* $R$?
**T** In a special case $R$ may be an object. But generally $R$ may include agents too, those not included in $A$. Control usually effects only *some* agents, and it is these agents that I include in $A$. The actions and other representations that constitute the inner mechanics of the system $S$ would be typically preserved and possibly somewhat modified.

The arrows in the figure indicate causal relationships. The controlling system $C'$ includes an agent $A'$ which restricts the freedom of the agent(s) $A$ in $S$. It also includes another agent, $R'$, which we call a representation of $S$. Its actions are restricted by $S$, while $R'$ itself restricts $A'$. Thus the causal loop is closed: a phenomenon known as *feedback*. However, the relation between $S$ and $C'$ is not symmetric; $C'$ controls $S$, but $S$ does not control $C'$. The effect of $A'$ on $S$ is direct and may be of any kind and degree, including a complete destruction (though, as I said, this is not the most interesting case). But $S$ effects $A'$ only through $R'$, which serves as a kind of transducer, or filter. The effect of $S$ on $A'$ cannot be greater than allowed by the changing states of $R'$.

In the most fundamental example of the control scheme, $C'$ is an organism, and $S$ its environment; the link $A' \rightarrow S$ represents the organism's actions, $S \rightarrow R'$ the sensation creating a certain perception of the environment by the organism, and $R' \rightarrow A'$ the decision taking by the organism on the basis of its perception of the environment.

The concept of a computing machine is a direct expression of control. The machine works on certain objects: input, intermediate and output.

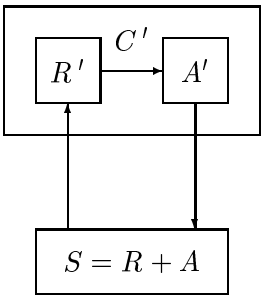


Figure 3: The control scheme

It is in a complete control, changing them directly. The inverse influence of data on the machine occurs only at certain moments and is limited to causing the machine to take one of a few possible ways of computation when executing conditional statements. The conditions in such statements create representations of the data field. For example, when the statement **if** x¡y **then**... etc. is executed, the representation $R'$ is the value of the Boolean expression $x < y$.

Control is often called for keeping some variable $x$ in a system around its desired value, or, more generally, achieving some goal expressible as a representation. Then we have a regulation scheme, or a scheme of purposive behavior, see Fig.4. In addition to the representation of the changing environment, the system $C'$ has one more subsystem which represents a goal, such as the ideal desirable value of $x = x_0$. In this special case $A'$ operates in this manner: whenever $x > x_0$, it performs an action which decreases $x$; when $x < x_0$, it performs an action which increases $x$. Generally, it compares the current situation $R'$ with the goal $G'$, and performs an action which makes the two closer.

**S** Did you say that the relation of control is antisymmetric?

**T** No. Only that it is not symmetric. It is possible that $X$ controls $Y$ and $Y$, at the same time, controls $X$. The control here may be destructive, then we call it a *conflict*, as when two men are fighting. Each one is effected by two channels: maintaining one's perception of the other, and taking his blows. Or the parties may mutually strengthen and build each other. This *circular control* we find between DNA and proteins in the work of biological

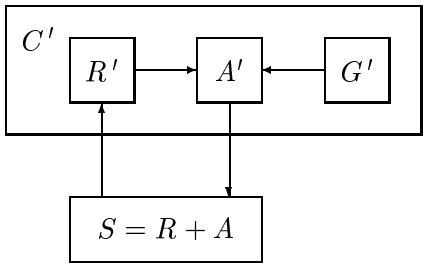


Figure 4: The scheme of regulation, or purposive behavior

machinery.

More interesting than conflicts are situations when hierarchies of control emerge. A metasystem transition is the emergence of a new level of control, usually accompanied by integration of a number of the pre-existing systems. In Fig.3 we call a metasystem the system which includes $C'$ and $S$; symbolically, $S' = C' * S$. If $C'$ exercises control over $n$ systems of the type of $S$, we can put this symbolically as $C' * (S_1 + S_2 + \ldots + S_n)$. A metasystem transition is a transition from a system to a metasystem:

$$S \ \rightarrow \ S' = C * (S_1 + S_2 + \ldots + S_n)$$

It was represented graphically in Fig.2.

**S** I am not certain what you mean by *transition*. Is it a physical process, or just a mental construction, as when we say: 'Consider a metasystem...'?

**T** All processes are, in the last analysis, physical processes. Yes, an MST is a physical process, an *action* of a special type which creates *a new agent*. We shall refer to such an action as an *emergence*.

Agents come into, and go out of, existence. One of the problems of philosophy has always been: how to distinguish simple ('quantitative') changes from the cases where something really 'new' emerges. What does it mean to be 'new', to emerge? In our theory this intuitive notion is formalized as the coming into existence of a new agent. An action can lead to an emergence of new agents. Take, again, radioactive decay as an example. A neutron suddenly *chooses* to break down into a proton, electron and neutrino. Whatever agents were associated with the neutron (its kinetic energy, e.g.) do not exist anymore. New agents emerge, such as the interaction between the new-born proton and electron, or the kinetic energy of neutrino.

**S** Is then the decay of a neutron a metasystem transition?

**T** No. I did not say that *every* emergence is a metasystem transition. But I can give a simple example of an emergence which is. Consider the formation of a hydrogen molecule from two hydrogen atoms. Before this act we have two agents, the two atoms. After it we have a new agent: a molecule as a whole. It is not the same as two free independent atoms. The atoms are still there, but they are controlled by the molecule; more precisely, the protons are controlled by the common electron shell. A molecule exhibits actions which did not exist when the atoms were free: vibrations, rotations. We have to accept the fact that a new agent has been born, and that it results from the integration of the atoms.

**S** Can you show how what is going on in the molecule fits your definition of control?

**T** Sure. This is a clear case of regulation. Our metasystem consists of protons $S_1$ and $S_2$ integrated by the electron shell $C'$. Let us first treat the system classically, and consider, for simplicity, only one of the three spatial projections. The relevant representation $R'$ is the distance $x$ between the atoms. Nothing happens as long as the protons are at the equilibrium distance $x_0$. Suppose some external agent pushes the proton $S_1$, passing some momentum $p$. The co-ordinate of the proton starts changing. This will cause a force $F$ opposite in direction to the displacement of the proton and roughly proportional to it. This force will start stopping the proton by changing its momentum and finally reverse its movement. If the initial perturbation was not too big, oscillations will result, but the molecule will preserve its identity. The three causal links in the control scheme are obeying the following equations:

$$\begin{array}{ll} (A \to R') & dx/dt = p/m \\ (R' \to A') & F = -k(x - x_0) \\ (A' \to A) & dp/dt = F \end{array}$$

In a quantum-mechanical version you will see the same factors, but in a different mathematical formalism. Metasystem transition is a fundamental feature of the world at all levels, including the basic physical processes.

**S** Interesting. You represented as a control scheme the relation between the basic elements of classical mechanics: co-ordinate, impulse, and force. I expect that crystallization is also a metasystem transition according to your theory.

**T** Quite true. This illustrates one of the characteristics of metasystem transitions. There are two such general characteristics: the *scope* of a metasystem transition and its *scale of integration*. The scope of a metasystem

transition is the original system $S$ in the control scheme. The scope may be vastly different. It is an atom in the case of the hydrogen molecule; it is a human being when a human society is formed. The scale of integration is the number $n$ of integrated systems $S_i$. It is 2 in the case of a hydrogen molecule, and about $6 \cdot 10^{23}$ for a gram-atom of crystallizing matter.

There is also an important physical characteristic of control and metasystem transition which cannot be seen on the control scheme, exactly because it is physical, and not cybernetical. It is the *energy* of the controlling agent. To break down a hydrogen molecule in its ground state, an amount of energy is needed, which is known as the *binding energy* of the hydrogen atoms in the molecule. Different phenomena in nature are characterized by very different energy scales; thus the binding energy of nucleons in atomic nuclei is orders of magnitude greater than binding energies accounted for by electrons. Therefore, agents may be stronger of weaker. Moreover, we can formally treat two independent hydrogen atoms as a system of two atoms, a 'molecule', the binding energy of which is zero. The concept of an agent has a quantitative aspect. An agent may be 'very weak', practically non-existent. But we abstract from this aspect when we draw control schemes.

Metasystem transitions may originate either 'naturally', or as a result of human activities. I put the reference to nature in quotes, because human activities are no less natural than processes we call natural – this is the whole point of the MST theory; we see both biological and cultural evolution as forwarded by metasystem transitions. The reason metasystem transitions take place in biological evolution is that they enhance survivability. In cultural evolution the most radical creative feats are also metasystem transitions.

I believe, David Hilbert was the first to use the prefix *meta* (from the Greek *over*) in the sense we use it in *metalanguage*, *metatheory*, and now *metasystem*. He introduced the term *metamathematics* to denote a mathematical theory of mathematical proof. In terms of our control scheme, $A$ in metamathematics is a mathematician who proves theorems (mathematical texts in natural language can be seen as $R$ in the scheme); $R'$ is a representation of mathematical texts in the language of formal logic; $A'$ is a metamathematician who translates texts into this formal language and directs and controls the work of $A$ by checking the validity of his proofs and, possibly, mechanically generating proofs in a computer. The emergence of the metamathematician is a metasystem transition. Complete formalization of the actions by $A'$ makes it possible to make the next metasystem transition, where $A''$ proves statements describing the activities of $A'$, in particular,

finding out that there are certain statements which can be neither proven nor refuted. We see here a three-level system: $A$ can be associated with the name of Euclid, $A'$ Hilbert, and $A''$ Gödel.

**S** A metasystem stairway.

**T** Exactly. The link from representation to action in a control scheme may or may not be semantic. In our example of the hydrogen molecule this link was not semantic, but a direct and unchangeable manifestation of a natural law: $F = -k(x - x_0)$. There is a class of metasystem transitions, though, which, by definition, is based on a semantic relation: this is when we embark on the task of *description*. Take an example from computer programming. Suppose you have a function $F(x)$. This means that you have a machine, i.e. agent, $F$, and various objects which may become values of $x$. Then you describe $F$ in a certain formal language $L$, which is understood by an interpreter $Int$. This description is a code; let us denote it as $code(F)$. To say that $Int$ understands $L$ is to say that $Int$ and $code(F)$ are in a semantic relationship: $Int$'s action using $code(F)$ imitates the action of $F$. A new control structure is created. Look, please, at Fig.3 to see how it fits the control scheme. The lower level $S = Ob + A$ is what we have at the beginning. Here $A$ is the machine $F$, $Ob$ is the collection of input and output objects for $F$. The representation $R'$ is the code $code(F)$. The agent $A'$ is the machine $Int$. The step from $F$ to $Int$ is a metasystem transition; $Int$ is a *metamachine*. In the paper by Glück and Klimov [2] you can find some examples of the use of the MST concept in computer science and mathematics.

To finish with the definition of metasystem transition, I should mention a special case where the control system $C'$ does not actually include, or use, representation $R'$. In other words, the actions of $A'$ are completely blind, chaotic. Such, apparently, is the control of cosmic rays over genes when they cause mutations.

## 3 Evolution

**S** I liked very much our dip in the waters of the Small Sebago lake.

**T** So did I. Sometimes I have my doubts about abstract thought versus concrete enjoyment. But let us try to have the best from both.

I want to discuss evolution now. According to the neo-Darwinist view, evolution takes place due to creation of random combinations of matter, with the subsequent struggle for existence, as a result of which some combinations

survive and proliferate, while other perish. Popper [9] describes this as the work of the general *method of trial and error-elimination.* Campbell [1] uses the term *blind variation and selective retention.* I speak simply of the *trial and error method.* I would rather not use the term 'blind', because in cultural evolution we often have informed and guided choices. But even with regard to biological evolution we cannot be sure, much less prove, that the variation is blind. It is true that we build our theory and check it against facts in the assumption that variations are blind. But the success of such a theory only proves that blindness is, sometimes, sufficient, but does not prove it is necessary.

¿From the viewpoint of a physicist, evolution can be seen as a search for *stability.* Consider a system of atoms in the framework of non-relativistic quantum mechanics and statistics. Its *configuration* is a set of values of all generalized co-ordinates of the system, e.g., the co-ordinates of all atomic nuclei and electrons. The system is described by Schrödinger's equation which includes the potential energy of the system as a function of its configuration. This function can be visualized as a distribution in the $n$-dimensional configuration space, where $n$ is the number of the degrees of freedom. The solutions of the Schrödinger equation for a given total energy $E$ of the system constitute the total set of all possible quantum states of the system. When we deal with a macroscopic system, though, it is never completely isolated, but constantly exchanges energy with its environment, so there is an interval $\Delta E$ of energy within which the exact energy of the system varies. Let the total set of states with energies around $E$ and inside $\Delta E$ be $S$. The system jumps randomly within this set from one state to another. The logarithm of the number of states in $S$ is the *entropy* of the system: $\eta = \ln |S|$.

We also can see the state of a macroscopic system in the quasi-classical approximation as a point moving in its *phase space.* The phase space has twice as many co-ordinate axes as the configuration space: for each generalized co-ordinate it includes the corresponding generalized impulse. The states available for the system make a surface of a given energy $E$, or, more precisely, the space between the surfaces for $E$ and $E + \Delta E$. The volume of this space measured in the units of the Plank's constant $h$ is the same number of quantum states $|S|$ as above. It defines the entropy of the system.

A quantum system tends to find itself in a configuration which has the minimum of potential energy. But the potential energy of a macroscopic system is an extremely complex and irregular function. It has a stupefying combinatorial number of local minima and maxima. If the system finds itself in a local minimum of energy, it must overcome a potential barrier in order

to leap to another minimum. The probability of such a leap includes the factor $e^{-b/T}$, where $b$ is the height of the barrier, and $T$ is the temperature of the system, i.e. the average energy per one degree of freedom. Hence if the barrier is much greater than $T$, the probability of jumping over it is exceedingly small.

Imagine a potential energy function which looks like a crater on the moon: an area $C_1$ surrounded by a pretty high (as compared with the temperature $T$) circular ridge. Let the phase space volume corresponding to $C_1$ be $S_1$. It is a subset of the total set of states $S$, so $S_1 \subset S$. Accordingly, as long as the system stays in $S_1$ its entropy $\eta_1$ is less than for the system free to be found in any state of $S$: $\eta_1 < \eta$.

Now the following fundamental fact is in order: given two quantum states, $s_1$ and $s_2$, the probability of transition from $s_1$ to $s_2$ is the same as that of the inverse transition from $s_2$ to $s_1$. If at some time the system is found in the state $s_1 \in S_1$, it may stay there until a transition to some state $s_2$ occurs, which is not in $S_1$: $s_2 \in S_2 = (S - S_1)$. But the probability that it will get back from $S_2$ to $S_1$ is even much less; for macroscopic phenomena it is so small that it is, in fact, impossible. Indeed, let the probability rate of a transition between the states of $S_1$ and $S_2$ be of the order of magnitude $p$. Then the probability of jumping from (any state in) $S_1$ to (any state in) $S_2$ is $p|S_2|$, while the probability of the inverse transition is $p|S_1|$. Recall now that $S_1$ results from a certain constraint on $S$. The properties of combinatorial numbers are such that if a constraint is removed, the number of combinations increases at a mind-boggling rate. Thus $|S|$ is not just greater than $|S_1|$, but many, many times greater. Hence $|S_2|$ is also many times greater than $|S_1|$. The probability of returning to $S_1$ will be less than that of escaping from it by the factor:

$$|S_1|/|S_2| \approx |S_1|/|S| = e^{-(\eta - \eta_1)}$$

For macroscopic phenomena the difference between the entropies will be macroscopic, and the exponent vanishing.

Hence the law of the growth of entropy. A system will not jump from the larger set $S$ to a smaller set $S_1$. When a system changes its macroscopic state, its entropy can only increase.

In this light let us look at stability. As long as the system stays in the state $S_1$, it preserves its identity. But sooner or later, under the influence of cosmic radiation, or just an especially big fluctuation of thermal energy, a quantum leap takes place and the system is in $S_2$. Some part of its

*organization*, defined as compliance with some specified constraint, is lost. The entropy went up. How can we bring the system back to $S_1$?

The answer is: we need a certain amount of energy to overcome potential barriers. But there is an additional requirement to that energy: it must belong to a single agent, or, maybe, to a very few agents. An agent in this context is a force or forces associated with one degree of freedom, or a few degrees of freedom, between which there is a strong interaction (note: even deterministic classical mechanics cannot do without speaking of *freedom*). A big system of atoms can be divided into some regions within which there is a significant interaction, while the interaction between the regions is weaker by orders of magnitude. The potential barriers of which we speak are regional. So are the corresponding degrees of freedom (generalized co-ordinates). A jump over a barrier changes equilibrium values of a few generalized co-ordinates. To make this jump the system must obtain an amount of energy comparable with the height of the barrier and concentrated on the co-ordinates which take part in the jump. In the language of agents, we need to pass to the agent of the jump the necessary amount of energy. Then it becomes possible to make a jump which amends the deteriorated organization, or creates it anew.

Given an amount of energy, we must ask an important question about it: is this energy concentrated on a single agent, or pulverized among a huge amount of nearly independent agents. The latter is thermal energy; the former is known in thermodynamics as *free energy* (freedom again!). It is only free energy which creates organization. Energy distributed between a great number of independent agents is useless for organization, because there is no force which could collect it into an amount sufficient for overcoming potential barriers, while the probability that this happens by chance is virtually non-existent.

So, we used a quantum of energy to overcome a potential barrier and create a desired regional configuration of atoms. When the point representing a configuration jumps from one side of the barrier to the other, its level of energy changes little, if at all. Then where does the energy we passed to the system go? In the last analysis, it dissipates between all the agents in the system, i.e. converts to the thermal form. If we want to have a stable system, such as a living system, there must be a way to get rid of this thermal energy, otherwise the temperature will raise higher and higher until rampant agents of thermal motion kill all organization around.

We come to the conclusion that if we want to see a stable or growing organization, there must be a relatively small number of agents which main-

tain the organization, passing to it in the process some energy, which later escapes the system as thermal energy. This flow of energy where it enters the system in a low-entropy form, i.e. vested in a few number of agents, and leaves it in a high-entropy thermal form, is essential for preserving organization.

**S** Interesting. I knew about the phenomenon of a potential barrier, of course, but I did not have a clear picture of how it is related to stability. I thought, the lower is the energy of a system, the more stable it is.

**T** This is not so. A system may be in a state with relatively low energy but surrounded by low potential barrier. It will have much greater probability to jump someplace than the same system in a state with a higher equilibrium energy, but surrounded by a high potential barrier. Stability is a feature of the configuration-energy function of the system. Theoretically, it is this function that is studied by cyberneticians and biologists. We are interested in the structure of the energy function of systems, the existence of local minima well protected by potential barriers. Evolution, and life itself, is the wandering of the system around local minima, in search of more and more protected configurations.

**S** I hope you are not a reductionist. I hope you do not think that cybernetics and biology can be reduced to a branch of physics.

**T** No. when we say that the reduction is possible 'theoretically', we simply indicate the place in our theory where physics borders with other fields of science. The energy function for a system of many atoms is an object of mind-boggling complexity. What can be done by methods of physics must be done, but this will never be enough.

Now let us think about possible paths to the points of stability. Given as the starting point an unorganized matter, say, the primary soup of various organic molecules, and a stable configuration as the desired end, one path will require climbing a mountain and then descending to the valley, while another path may lead around such mountains and overcoming much lower barriers. The role of enzymes in biological processes is to make such paths possible that reach the goal with minimal supply of energy.

Consider the question of *damage control*: when a system deviates from stability by jumping over the surrounding potential barrier, how is it brought back? We discussed it above in terms of quantum mechanics. We came to the conclusion that an agent is necessary to perform a reverse transition. If such an agent is available, and if it is automatically called as a result of the disorganizing transition, I will call such damage control, and the stability it achieves, *causal*. Causal stability, in fact, is familiar in the context of

macroscopic, classical (not quantum) description. When we speak about the stability achieved through control and regulation, we speak about causal stability: all links in the control scheme are causal relationships. This is how stability is achieved in the macro world.

When we deal with phenomena on the level of individual atoms and molecules, causal stability becomes difficult to organize. The number of possible quantum transitions is huge, and to tie to each of these transitions a corrective agent – in the world where only probabilities are predictable – this is some task! This is the case where it is easier to make a toy from scratch according to its description than to fix it. And this is another method of damage control and stability, *replication stability*, which is the main method of achieving stability in the living systems on the atomic level.

To have a description is the key element. This role is played by molecules of nucleic acids DNA and RNA. Proteins are responsible for creation of a huge variety of configurations. There is also a universal agent which carries energy necessary to jump over potential barriers. This is the molecule of adenosine triphosphate (ATP). The relation between these three basic elements of life is that of *circular control*, shown in Fig.5. Proteins are synthesized according to the DNA/RNA code; but it is the spatial patterns of protein enzymes that determine where the agent ATP will do its job and restructure a configuration. In particular, this makes the process of replication of nucleic acids possible.

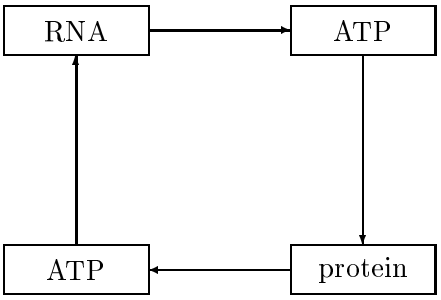


Figure 5: Circular control at the molecular level

At the molecular level life relies on replication for stability. All living structures die because of the relentless entropy. But before they die they produce modified copies that survive.

Combination of replication with the trial and error method makes metasystem transition the main vehicle of evolution. Integration of substructures

at all levels of evolution creates, generally, more chances for stability because of two very general factors. The first is the simple *geometric* factor: the binding energy is proportional to the volume of the structure, while the external perturbation is proportional to its surface. The second is the *probabilistic* factor: the more systems are integrated, the greater is the probability that something useful will be discovered by the method of trial and error, simply because there is more of trial.

There is, however, one more factor, *combinatorial*, which works in the opposite direction. Suppose that you have some number $n$ of structural units and you expect that there is an arrangement (loosely speaking, a combination) of these units which will be significantly more stable than the other combinations. Using the trial and error method, you test one arrangement after another and check their stability. The number of possible arrangements grows catastrophically with the number of units $n$. Let us take the simplest case where the units always form a linear structure. Then our arrangements are permutations of $n$ elements, and there are $n!$ of them. If $n = 5$, the number of arrangements is 120. You can easily try them all. If $n = 100$, there are some $10^5 6$ arrangements. Even if there are billions of billions of useful arrangements of 100 units, the probability that you will find one in 10 billion years, is vanishing. Neither can mother nature try all possible arrangements of many units. Instead she tries the arrangements of a relatively small number of units, achieves partial success in terms of stability, and then makes a metasystem transition to integrate any – possibly very big – number of successful arrangements into a metasystem. The metasystem thus emerged becomes a new structural unit for further combinations and arrangements. Such is the origin of hierarchical structures in evolution, both biological and post-biological. This is a way to use the geometric and probabilistic factors, but avoid combinatorial explosion.

**S** How does the new control level emerge? There is something mystical about it. I cannot see the mechanics of it.

**T** OK, let us speak about the mechanics of metasystem transitions. Apparently, there is no general method for it; the mechanics depends on the physical nature of the system. But there are some common features of the process. A new agent can emerge as the result of the collective effect; this is, in Hegelian and Marxian terminology the transformation of quantity to quality. Crystallization is an example. As long as there are only two or three atoms, there is no crystal, even if they make a right configuration. A certain minimum is necessary for stability; then the process of crystallization starts.

A different mechanism of emergence we see in the control structure of

animal and human groups and societies. A kind of instability with respect to control often exists. An animal who happens to be a little stronger than others becomes dominant and takes more and better food. As a result it becomes even stronger an more domineering. In a human group, a member who happens to be the first to have a gun can prevent others from having guns. On this principle multilevel hierarchies in human societies have come into existence.

There is a general feature of metasystem transitions, which I call in [12] the law of *branching growth of the penultimate level* of control. Initially, the integration of replicated subsystems can take place only on a small scale, because of the combinatorial factor we have discussed. However, when the needed combination is found, and a new controlling agent has emerged, it becomes, typically, possible to control greater and greater numbers of integrated subsystems, and this is advantageous for stability because of the geometric and combinatorial factors. An integration on a grand scale starts. The emergent agent is on the ultimate control level of the emergent system; the integrated subsystem make up the penultimate level. A metasystem transition leads to multiplication of these penultimate-level subsystems. When nature had discovered the principle of coding protein forms with sequences of four nucleotides, the growth of the number of nucleotides began, resulting in huge molecules with many thousands of nucleotides. When the concept of a cell that can cooperate with other cells emerged, multicellular organisms started being formed with growing numbers of integrated cells, until they reached the sizes of present day animals. The same with human society. A well organized society starts growing exponentially. All these are instances of a general cybernetic law.

**S** You speak of evolution as if it were conceived and realized by what you call 'nature', similar to how we ourselves design things and construct them.

**T** You should take this as a metaphor. Actually, all my pronouncements of that kind can be translated into the standard language of trial and error, or blind variation and selective retention. But I would not be quite honest if I did not tell you that sometimes I think that there is more behind that metaphor than we are ready to accept at the present time.

**S** I feel a closet vitalist in you.

**T** Vitalist or not, I said all I could say, at the present time, about the evolution at the molecular level. Let me turn to the macroscopic level and causal damage control, which will ultimately bring us to the focus of our discussion: supreme human values and the future of the world.

Repeated metasystem transitions create control hierarchies. One verti-

cal cell of such a hierarchy is pictured in Fig.6. Here the control unit of the second level $C''$ controls the control unit $C'$, and may be controlled, in its turn, by a higher level unit. Two flows of level-to-level information are formed when control units are connected vertically. Creation of representations proceeds upwards: $R'$ is a representation of the ultimate object of control (environment) $R$, while $R''$ is a representation of the representation $R'$, etc. Execution of goal-directed actions proceeds from top down. Every control hierarchy has its top level $C^{(t)}$. At this level, the representation $R^{(t)}$ is the most advanced abstract representation in existence, and the goal $G^{(t)}$ is the supreme goal: survival and proliferation in the case of an animal.

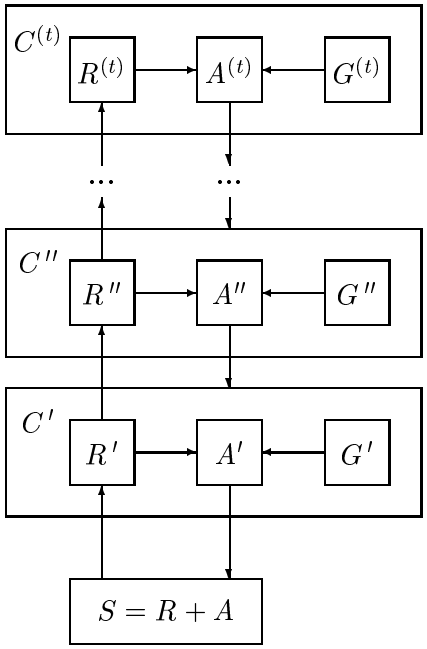


Figure 6: Hierarchical control, a vertical section

Notice that the growth of control hierarchy adds new loops of feedback without destroying the already existing loops. The action $A''$ informed by (i.e. trying to achieve) the goal $G''$ controls the agents $A'$, not the agents $A$ of the environment. $A'$ still remains in the immediate control of the environment and it tries to achieve its own goal $G'$, working in the feedback loop $A'(A + R)R'A'$. The second level agent $A''$ controls, of course, the whole first-level system $C'$, including $G'$. Thus $G'$

is nothing but a subgoal of $G''$. The latter can be seen as a program which calls the former as a subprogram. When the goal is achieved, control returns to $A''$ which can now set another goal $G'$, as defined by $G''$, and let $A'$ to achieve it. So it works by loops within loops: a situation familiar from computer programming.

Fig.6 shows only a vertical section of the control hierarchy. In reality $C''$ controls not one system $C'$ but many similar such systems, which come into being by integration in a metasystem transition (see Fig.2). For example, when a light receptor emerges and is used to control movement, it becomes possible for an animal to have more of such receptors to receive more of potentially useful information. But to make this information really useful, a new level of control must emerge with a representation which compresses information and translates it into action. When a control mechanism of that kind emerges, the number of light receptors grows rapidly by the law of the penultimate level. In the hierarchy of goals we also see integration of subsystems: $G''$ may include many subgoals $G'$.

So, a better idea of a control hierarchy will be given if we replace each control unit by many units all connected to the unit above, as in Fig.2. The control hierarchies which actually emerge in evolution are not, of course, so regular and tidy. Remember that the *scope* of a metasystem transition may vary; this jump may occur in different subsystems at different levels.

I set aside the question of how the relation between $R'$ and $A'$ is settled down. This relation may be dictated by a law of nature, as is the case for the hydrogen molecule. But in biological systems this relation is *semantic*, which means that the agent which makes the transition from $R'$ to $A'$ is formed of an object and its interpreter. In a rough model, this object can be seen as a table of pairs $(R'A')$, with the interpreter working as we do when we use tables: given an $R'$ it looks through the pairs, finds (hopefully!) a pair with $R'$ as the first component, and activates the $A$ which is the second component of the pair. This is the classical concept of a function in mathematics: the function of behavior.

Semantic relations, unlike relations reflecting a specific law of nature, provide to the user the freedom of choosing and varying the table that defines the relation. This is an asset in the struggle for existence, because it gives the system the freedom to evolve to greater stability, by fixing the table in a certain fashion. Specifically, it works according to the principle of trial and error: the table varies randomly, and the creatures where the table is 'wrong' become extinct, while among the living creatures we find only those with 'right' tables.

**S** So, the relation then becomes defined because the table becomes defined, and the freedom to have various tables is lost.
**T** Yes, this is the paradox of freedom in life and evolution. Freedom is needed in order to make a choice, thereby loosing the freedom. You consume freedom, like engine consumes fuel. This is the way to survive.

You could have noticed that I am not trying to present any structural scheme or model of how the behavioral function works. This is not my piece of cake, and anyway, at the present time we know so little about actual physical mechanisms of the behavior of animals, that not much could be added, probably. My method of analyzing the course of evolution has been purely *functional*. But I contend that even staying on this high level of abstraction, it is possible to come to quite definite conclusions. This was, in fact, the main idea of em The Phenomenon of Science.

If we refrain from drawing pictures of internal states and processes, what is left for us is only to use the principle that evolution proceeds by metasystem transition. In the most abstract way metasystem transition can be representing by the formula:

$$A' \ = \ \text{control of } A$$

In terms of such metasystem transitions, I can trace the major stages of the evolution of life on the Earth. I start from the unicellular and the most primitive multicellular animals, like Coelenterata.
**S** But the leaving cell is already an extremely complex machine. Life does not starts with the cell. It starts with the first macromolecules.
**T** I agree. It would be very interesting to trace the metasystem stairway down into the history of life from macromolecules to the cell. My knowledge of molecular biology is not enough for even starting. Maybe somebody will do this later. But for evolution from the cell to the human society, I believe, the layman's knowledge of relevant subjects allows us to put up a fairly convincing picture.

So, we start with the most primitive animals. Unlike plants, they have an apparatus that allows them to take actions of their own and to control those actions through irritation of nerve cells. Take a hydra. It has two layers of cells containing muscle fibers which contract when irritated, and nerve cells (receptors) which can irritate and pass irritation to muscle fibers. If a hydra is pricked with a needle it squeezes itself into a tiny ball. The emergence of this apparatus is a metasystem transition (MST) from the stage of primitive plants where there are no self-induced actions. This MST

is defined by the formula:

$$\text{cell-irritation} \;=\; \text{control of actions}$$

Indeed, we see in the hydra all elements of the control scheme in Fig.3. $R'$ is represented by the receptors in the ectoderm which get irritated and pass irritation to muscle fibers; $A'$ stands for the muscle fibers, which act when irritated.

The next MST in functional terms becomes possible due to the structural MST of great importance: integration of more and more cells into big multicellular organisms. Integration is accompanied by specialization. In particular, specialized nerve cells emerge. They make up a complicated nerve net where one cell stimulates (irritates) another, and signals from receptors may pass a long and tortuous way through the network before the effector cells are stimulated and trigger bodily effects in the organism. We call the whole process a (complex) *reflex*. The development and perfection of biological nerve nets takes place, probably, by a series of metasystem transitions, but thinking in functional terms we can unite all them into one with the formula:

$$\text{reflex} \;=\; \text{control of cell irritation}$$

The final stimulation of effectors does not immediately follow irritation of a receptor, but is the result of the work of a complex network which controls irritation of millions of nerve cells. The control becomes hierarchical, as in Fig.6, with the flows of representation and action in opposite directions.

If the preceding stage of evolution can be called the stage of hydra, the new stage can be characterized as the stage of ant. Reflexes and behavioral programs are hierarchical and complicated, but they are defined at birth and do not depend on individual experience.

What is the next stage? We can call it the stage of dog. Let us think of the reflex as a set of $R'A'$ pairs. Unlike the ant (or at least the schematic ant according to our definition) the dog can make an association between a situation represented by $R'$ and an action $A'$ which happens to enhance the dog's viability. I hypothesize that at such a moment the dog feels a positive emotion, and that an emotion, generally, is an internal view of the organism's action which enhances viability. The dog is capable of *learning*. Experiencing a few times, or even once, the emotion of a successful response $A'$ to the situation $R'$, it makes the association $R'A'$ and keeps it in its memory.

So, the formula of this MST is:

$$\text{association} \;=\; \text{control of reflexes}$$

Here by association I mean the *action* of association; associations as elements of the table exist already at the stage of ant.

At the stage of dog a new phenomenon emerges: the phenomenon of modeling the environment. It is natural to assume that if the dog can associate a situation with the immediately following action, then it also can associate such pairs following each another. Thus if in the flow of perception and action after the pair $R'_1A'_1$, there follows the pair $R'_2A'_2$ then the sequence $R'_1A'_1R'_2A'_2$ may stay in memory. But this sequence can also be seen as $(R'_1A'_1R'_2)A'_2$, i.e. as consisting of the triplet $(R'_1A'_1R'_2)$ followed by the action $A'_2$ (there is a hidden assumption here that the code in which information is stored is associative, like when it is stored in chains of symbols, which seems natural for living matter ever since the linear structure of chromosomes was discovered). This triplet is a *model* of the world, as it was defined in the epistemological part of our discussion (see Fig.1). Those triplets which predict false result $R'_2$ of acting $A'_1$ in the situation $R'_1$ have slim chances to bring positive emotion with any subsequent $A'_2$ – it may be too late. So, such triplets will not enter memory. Those triplets making correct predictions allow to take two correct actions; they stay in memory and constitute the animal's *knowledge*. I started philosophizing with the definition of knowledge as a model of the world. Now we see why knowledge thus defined emerges in the course of evolution.

**S** I suspect that the next MST is in order, and that this stage of evolution will be characterized as the stage of man.

**T** You suspect right. The functional formula of this transition is:

$$\text{thinking} \;=\; \text{control of associations}$$

It remains to demonstrate that the features of human thinking, as we know it, do indeed fit this formula. I will try to do this tomorrow, if you do not mind. It is time to have a good dinner.

# 4 The Human Being

**T** Good morning. Let me start with a commonly used disclaimer. We still know so little about the process of thinking that any theory claiming to

explain the essence of this phenomenon is hypothetical. Thus my conception of thinking must also be treated as a hypothesis. However, it indicates the place of thinking in the row of natural phenomena and gives a coherent and consistent explanation of the observable manifestation of thinking. I try to avoid any assumptions regarding the concrete structure and working mechanism of the human brain, except that there are some structures in it that allow for the new level of control – control of associations. I reason mostly in functional and phenomenological terms.

**S** Do you speak of thinking as a phenomenon, or specifically about human thinking. After all, animals think too.

**T** I do not know what is thinking as a phenomenon, if it is not human thinking. Yes, in a certain sense higher animals do think, but I stress that I speak of *human* thinking, having in mind those feature which are developed in full only in the human being, even though they may be present in a rudimentary form in some animals.

**S** All right. So, what is control of associations?

**T** Remember we spoke yesterday of association triplets $R_1A_1R_2$ etc. At the stage of dog they arise spontaneously and are selected when they are beneficial. At the stage of man these triplets become objects of work for the next level of control. In a metasystem transition some things that were once fixed and determined from birth, or by external factors, become variable and subject to the trial and error method. Control of associations, like every metasystem transition, is a revolutionary step directed against slavish obedience by the organism to environmental dictatorship. As is always true in the trial and error method, only a small proportion of the arbitrary associations prove useful and are reinforced, but these are associations which could not have arisen directly by chance or under the influence of the environment. A dog may be trained to drag a bench to a fence, climb up on the bench, and jump from it over the fence. But if the dog was not taught this, it will not figure it out with its own mind, even though it may know how to drag the bench and how to jump from it over the fence.

**S** You do not say so. I had ...

**T** Please, don't. I know. You may have had an exceptionally clever pet, but this is beside the point. We discuss two types of brain and behavior, not the abilities of concrete creatures. I speak of a schematic dog, and take the simplest acts of human thinking, which are on the verge of what a dog can do. You cannot deny that there is a huge gap between a man and a dog. I would not believe if you told me that your dog solved differential equations.

Back to *our* dog, it has the model (triplet) $R_1A_1R_2$ where $R_1$ is the

bench off the fence, $A_1$ is dragging the bench, and $R_2$ is the bench at the fence. It also has the model $R_2A_2R_3$ where $A_2$ is jumping on the bench and form bench over the fence, and $R_3$ is the desired result (suppose the food is behind the fence). However, it cannot combine these two models into $R_1A_1A_2R_3$, and without this action the problem is not solvable. The dog does not do $A_1$ because there is no reward for it, and cannot do $A_2$ (which promises a reward) because the situation $R_1$ does not allow it. We, humans, can combine representations in our mind, and we call it *imagination*. The schematic dog lacks imagination which is necessary to solve the problem.

**S** Wait a moment. You define imagination as putting two associations together, and state that this becomes possible with a metasystem transition. But you had it already at the stage of dog, when you combined $R_1A_1$ and $R_2A_3$. Then where is the metasystem transition? What is new?

**T** The combination of representations into models is not new. It is there at the stage of dog already. New is the mechanism of the combination of representations and the closing of the feed-back loop. At the stage of dog spontaneous and externally defined combinations of representations are translated into actual behavior of the animal, and its survival is at stake in the process of selection of correct models. At the human level the whole cycle of the trial and error takes place inside the brain, in our imagination. We discard the sequences of actions which lead to undesirable situations, without actually doing these actions in real life. This is a new level of the control of associations. Animals *have* models of reality. Humans *create* them.

The advantages of this metasystem transition are enormous. First, the trial-and-error method inside the brain works many times faster than it works when the evaluation of a situation takes place in real life. Second, when trial and error occurs only in imagination, then situations where severe damage is inflicted on the organism or its death becomes imminent can be recognized as such, without actually being experienced. This is, obviously, a great advantage in the struggle for existence.

**S** All right, imagination is accepted as control of associations. What else?

**T** *The use and manufacture of tools.* This is usually indicated as the first decisive difference between humans and animals when speaking of the origins of human beings. The borderline here lies between using tools and making them. Animals use tools occasionally, and sometimes very skilfully. But making tools requires imagination, and this is a human privilege.

**S** But a chimpanzee can manufacture a stick in order to extract a banana from a tube.

**T** This is, again, a border case. But show me a chimp who can make a stone axe, and I will say that this is a primitive man.

There is another form of behavior which is a harbinger of the coming metasystem transition. It is *play*. I am not referring to behavior related to mating, which is also often called play, but rather to 'pure' and, by appearance, completely purposeless play – play for pleasure. This is how the young of almost all mammals play with one another, or how a cat plays with a piece of crumpled paper on a string. Play is usually explained as a result of the need to exercise the muscles and nervous system, and it certainly is useful for this purpose. But how this behavior becomes possible? The playing cat is not deceived into thinking that the paper is edible. Its representation of the paper is not included into the concept 'prey'. However, this representation partially activates the very same actions normally included into the concept 'prey'. Similarly, a wolf frolicking with another wolf does not take its playmate for an enemy but up to a certain point it behaves exactly as if it did. Play includes an arbitrary establishment of association between representations, such as a crumpled paper and a real prey. As a result there arises a new representation which, strictly speaking, has no equivalent in reality. We call it fantasy.

But let me move further. Unfortunately, I am not familiar with the contemporary theories of emotions. So I will base my exposition on my own simple, home-made theory. I believe that higher animals have positive emotions when the state they are in, and/or the actions they perform (remember, in my metaphysics this is the same) are favorable for the survival of the species; and they have negative emotions in the opposite case. If so, the emergence of the new apparatus of control over association of representations, which as we have seen, enhances survivability, must call to life new kind of emotions, which are characteristically and quintessentially human. These emotions are experienced when the purpose of the new apparatus – the creation of new models of reality – is successfully achieved. In the general form, I can call these emotions *the joy of revelation*; in particular, they include *the sense of the funny* and *the sense of the beautiful*, as well as the *religious feeling*. The corresponding negative emotion is nothing but *boredom*.

What makes us laugh? A disruption of the 'normal' course of events which is completely unexpected but at the same time natural, and in hindsight entirely understandable; an unexpected association, meaningless at the first glance but reflecting some deep-seated relationship among things. All this, of course, creates a new model of the world and gives pleasure propor-

tional to its novelty. When it is no longer new it is no longer funny. When someone tries to make us laugh using a very familiar model he only makes us bored. Another situation occurs when some people laugh, while another glances around uncomprehending. 'He did not get it' they would say. The joke was two subtle for that person; it relied on associations he did not have. The funny is always on the borderline between the commonplace and the unintelligible.

The sense of the beautiful is more subtle and mysterious than the sense of the funny. But here too we find the same dynamism related to the novelty of the impression. Too frequent repetition of a pleasing piece of music creates indifference to it, and finally revulsion. A sharp sensation of beautiful is short in duration; it includes an element of revelation, enchanted surprise. It can also be described as the sudden discernment of some deep order, correspondence, or meaning. In cybernetic terms, this is creation of a new model which uses some of our dormant associations as building blocks, of which we were not aware, and would not be aware if the artist did not reveal them to us through our own sense of beautiful. Like the funny, the beautiful is on the borderline between the commonplace and the unintelligible. A banal melody or a primitive geometric ornament will not elicit a sensation of the beautiful in us; we have already this model in our brain. But a Neanderthal man could, probably, be shaken to the depth of his soul upon seeing a series of precisely drawn concentric circles. The borderline at which we find art is shifting in the process of esthetic education. The attempt of some schools of thought to explain out the beautiful by reducing it to the narrowly understood useful, like Chernyshevski did, is pitiful. Pure esthetic education trains the brain to perform its highest and most subtle functions. The models created in the esthetic education must unquestionably influence the person's perception of the world and his creative ability. How exactly this happens is still unknown. Esthetic education is the more precious the less we know what we can substitute for it.

**S** It looks convincing. But I am especially interested to hear your interpretation of the religious feeling.

**T** That will come soon, but first I want to interpret the other two characteristically human features: *planning* and *overcoming instincts*.

Goals being elements of representations, the ability to associate representations arbitrarily means the ability to make plans arbitrarily. Man can decide as follows: first I will do $A$, then $B$, then $C$, and so forth. The corresponding chain of associations arises. He can decide that it is absolutely necessary to do $X$. The association '$X$ – necessary' arises. Action plans of

animals are always part of a more general (standing higher in the hierarchy) plan and, in the end, part of instinct. Instinct is the supreme judge of animal behavior – its absolute and immutable law. Man also inherits certain instincts, but thanks to his ability to control associations he can get around them and create plans not induced by instinct, and even contrary to it. Unlike animal, man sets his own goals. They partly come from his social environment, partly as the result of a free creative act.

**S** Why do you say 'control of associations' and not just 'control of representations'?

**T** That would be a stronger hypothesis, and I do not know if it is justified. Control of associations is a special kind of representation control where we are limited to combining and arranging some of the pre-existing representations only, but cannot create completely new representations from scratch. Can we imagine something that is not assembled from pieces we experienced in real life? I do not think so.

As for the religious feeling, my hypothesis is that it is the emotion corresponding to the setting of the supreme goal of the behavioral hierarchy. Different states and actions on the animal level, in animal or man, produce different emotions: the satisfaction of saturation is different from the satisfaction of sexual drive, even though they may be phenomena of the same kind. In the same way, the feeling of the beautiful which is triggered by the creation of a new model, is a different phenomenon than the feeling accompanying the setting of the supreme goal, even though these phenomena are of the same origin: control of association of representations.

When I speak of setting the supreme goal, I do not mean *any* goal, like getting rich, or becoming the president of the USA. That would be regular service goals from the middle of the goal hierarchy, even if the individual has no idea of higher values. Supreme goals in my understanding are characteristically human. They must include the realization of one's mortality and go beyond death, becoming supra-personal and somehow relating one's personality to eternity. The idea of Evolution on the cosmic scale also belongs to this category; it is, essentially, a religious idea, even though it is, at the same time, an established scientific theory.

Religious feeling belongs to the class of joys of revelation. When the supreme goal is set, it becomes clear to the individual what he or she must do. It becomes clear what is right and what is wrong. Apparently, this feeling is stronger when the person does not realize the free, arbitrary nature of the setting of the supreme goal. Then it is perceived as a discovery, revelation, God's blessing. But even if one comes to set the supreme goals as a free act,

religious feeling still is there.

**S** Well, this is much less convincing to me than your interpretation of the funny and beautiful. I doubt than a person who does not believe in God, or something like that, can have a real religious feeling. As we know from literature, a strong religious feeling often brings about an ecstasy.

**T** I think that these two phenomena should be treated separately. They are very different. Ecstasy, I believe must be treated as a physiological phenomenon, like dizziness, or pain.

**S** . Do *you*, not being a believer, have a religious feeling?

**T** When I first read, as a small boy, about evolution, the origin of species, and emergence of man, I was awed. I had a strongest feeling which I cannot call other than religious. It was very similar to how some people describe their childhood experience of entering church for the first time. I also can say that when I sorted out my ideas about the supreme goal, I did have a feeling which could be called religious, and something of it persists all along.

But let us go on with the consequences of the metasystem transition we are discussing. As long as we focus our attention at the separated human being, we cannot appreciate how revolutionary these consequences are. The frog is more intelligent than the jellyfish. The dog is more intelligent than the frog. The ape is more intelligent than the dog. Now there appears a creature that is more intelligent than the ape. So what?

The revolution that allows us to state that a new era in the evolution of the world starts, the era of reason, was made by the appearance of human society which possesses a definite culture, above all *language*. The creation of language by humans is a direct result of the metasystem transition to the control of associations. Once again we see that there is a borderline phenomenon: all social animals, including ants and bees, have languages for exchange of information. The difference between these and the human language is of the same kind as in the case of tools. The language of animals is instinctive; it develops as part of evolution of the species. But a human being *creates* a language by freely associating a name with its meaning. In a short biological time languages come into existence which contain hundreds of times as many different elements as animal languages, and allow their combination resulting in an infinite number of messages to send and to understand.

Language arises as a means of communication among members of a primitive community. But once it has arisen, it immediately becomes the source of other, completely new, possibilities which go beyond communication. It becomes a means of the creation of *new models of reality*, such models which

nature did not put into our heads. Doing arithmetic is the best example.

Imagine a primitive man who observes from his hiding place how members of a hostile tribe walk in and out of a cave. If three men come into the cave and two go out, he will know that one enemy is still in the cave: this is the work of a model which is built into his brain. But what if twenty enemies enter and nineteen exit? The brain model is of no use. But one can use fingers or pebbles or whatever is at hand to create a model of the enemies in the cave. The man will still use his brain models to perceive enemies as distinct objects in counting, but the representation of situations is now implemented in external material: fingers, pebbles, etc., not in the brain's stuff. If tool is a continuation of human hand, then language is a continuation of human brain.

**S** Do you use here the term *model* in the same sense as in the epistemological part of our discussion.

**T** Yes. I leave it to you to interpret counting in terms of that scheme.

There is an analogy between the emergence of language and the emergence of nervous system on a previous stage of evolution. Nerve cells also arise as means of interaction and co-ordination in the cell community, but once having arisen they develop into more and more complicated formations which serve the purpose of modeling the world. Finally, such a huge and wonderful instrument as the brain of mammals comes into being. Human language also develops into such a wonderful means of knowledge as the body of contemporary science, which is nothing but a huge linguistic model of the world.

The appearance of language signifies one more distinctive human feature: self-knowledge. The animal has no concept of itself; it does not need this concept to process information received from the outside. Its brain can be compared to a mirror that reflects the surrounding reality, but is not itself reflected in anything. In the most primitive human society each person is given a name. In this way a person, represented in the form of sentences containing the person's name, becomes an object of his or her own thought and study. Language is a kind of second mirror in which the entire world, including each individual, is reflected and in which each individual can see (in fact, cannot help but see!) his own self. The era of reason is the era of self-knowledge. The system of two mirrors, the brain and language, creates the possibility of a vast multitude of mutual reflections. This gives rise to the unsolvable riddles of self-knowledge, above all the riddle of death.

Control of associations, which is a metasystem transition in the structure of the brain – started another metasystem transition, *social integra-*

*tion*, the unification of human individuals into a who;e unit of a new type: human society. All human history has gone forward under the banner of social integration; relations among people have been growing qualitatively and quantitatively. This process is taking place at the present time, very intensively in fact, and no one can say for sure how far it will go.

Owing to the existence of language, human society differs fundamentally from animal communities. People have *contact by brain*. Language is not only a continuation of each individual brain but also a general, unitary continuation of the brains of all members of society. It is a collective model of reality on whose refinement all members of society are working, one that stores the experience of preceding generations.

This makes human society radically different from animal communities. We know communities of animals, such as ants, where individuals are so adapted to life within the community that they cannot live outside it. The anthill may be justifiably called a single organism; that is how far interaction between individuals and their specialization has gone in it. But there is no contact by brains; there is no creation of new models of reality. No fundamentally new possibilities are opened here by integration. This is not a new stage of evolution. The anthill freezes in its development: an evolutionary blind alley, apparently.

Society can be viewed as a single super-being. Its 'body' is the body of all people plus the objects that have been and are being made by people: clothing, dwellings, machines, books, etc. Its 'physiology' is the physiology of all people plus the *culture* of society, which I understand in a very wide sense as a certain method of controlling the physical component of the social body and the way that people think. The emergence and development of the social body marks the beginning of a new metasystem transition with the functional formula:

$$\text{culture} \;=\; \text{control of thinking}$$

**S** This does not agree with your all-inclusive definition of culture.

**T** Yes, to some extent. But I do not know how to separate from the all-inclusive culture that part which controls our thinking. Even the way you were taught to lace your boots controls, to a degree, your thinking – at least, with respect of boots.

Heylighen [4] expresses the view that social integration creates only a *supersystem*, not a real *metasystem*, because it brings "merely additional constraints on the exchange of thoughts, not on the development of new systems of thinking". But it is the society as a whole which converts thoughts

of individuals into systems of thinking and then implants these systems in the heads of subsequent generations. An isolated human individual could not create our culture, even if he was give million years for this task. The emergence of the social super-being is a large-scale MST which closely parallels an earlier MST: the emergence of multicellular organisms, and especially their nervous system. A single nerve cell cannot do much in terms of adjustment of behaviour to the changing environment. It is *interaction* between nervous cells that creates the world's models. The same with the human society. A supersystem transition, i.e. integration, is a necessary (though not sufficient: look at ants) condition for a metasystem transition. The two processes proceed in parallel.

The emergence of human society is a large-scale metasystem transition, in which the subsystems being integrated are whole organisms. It may be compared with the development of multicellular organisms from unicellular ones. But its

The emergence of the human super-being is even more significant than the emergence of multicellular organisms. If it is to be compared with something, it is only with the emergence of life itself. For the emergence of human society signifies the emergence of a new mechanism of evolution. Before it, the development and refinement of the highest level of organization, the brain device, occurred only as a result of the struggle for existence and natural selection. This was a slow process requiring the passage of many generations. In human society the development of language and culture is a result of creative efforts of its members. The selection of variants involved in the trial-and-error method now takes place in the human head. It becomes inseparable from the willed act of a human person. This process differs fundamentally from the process of natural selection in the genotype-phenotype cycles. It is incomparably faster. Cultural evolution takes over from biological evolution. The human being becomes the point of concentration of Cosmic Creativity.

Being human ourselves, we cannot look indifferently at the change from biological to cultural evolution, because cultural evolution depends on us; it is of our own making. I would like to quote from Teilhard de Chardin [11]:

"In fact I doubt whether there is a more decisive moment for a thinking being than when the scales fall from his eyes and he discovers that he is not an isolated unit lost in the cosmic solitudes, and realizes that a universal will to live converges and is hominized in him. In such a vision man is seen not as a static center of the world – as he for long believed himself to be

– but as the axis and leading shoot of evolution, which is something much finer."

**S** You obviously called your book, *The Phenomenon of Science* in parallel to Teilhard's book *The Phenomenon of Man.*

**T** Yes, and now I want to explain why. I see science as the apex of human culture. This is not an expression of my personal taste or love for science. I trace the evolution of human culture in the same terms as biological evolution, namely, as a sequence of metasystem transitions, and this sequence leads to science as the highest point in the hierarchy of control. Hence the development of science defines the future of the evolving Universe. I believe that the place and the role I ascribe to science is derived from an objective reality. So let me make a brisk run through the most visible metasystem transitions in the evolution of human culture.

I spoke of the manufacturing of tools as a typically human activity. There is an interesting detail. The difference between Upper Palaeolithic and Lower Palaeolithic periods is that *composite* implements appear, i.e. such that consist of two or more objects, e.g. a spear with a stone point. It may seem trivial to us, but it was not such for our ancestors. Indeed, the combination of two or more things into one whole which serves a definite function is a metasystem transition. Even in historical times a population was discovered, which did not know how to make composite tools. These are the indigenous inhabitants of Tasmania. They had the stone hand axe, sharp point, a crudely shaped cutting tool, two kinds of wooden clubs, wooden spear, stick and spade. But apparently, they did not have a single composite tool. They did not know how to attach a stone to a wooden handle.

The next metasystem transition is known as the *Neolithic revolution.* This was a transition from hunting and gathering to livestock herding and agriculture. The animal and plant worlds, which until that time had been only external, uncontrolled source of food, now became subject to active control by human beings. A typical MST. Riding horses, which had so great historical consequences, is also an MST, as well as plowing with oxen.

Suppose you know how to make a metasystem transition from a given system $S$, which belongs to some class $C$, to a metasystem $S'$. And suppose that $S'$ also belongs to the class $C$. Then you know how to make an MST from $S'$ to $S''$, then to $S'''$, etc. A *metasystem stairway* emerges, potentially infinite. I call the system that embraces all these systems and makes a growing metasystem stairway possible an *ultrametasystem.* When man learned how to make tools, and how to make tools for making better tools,

he created an ultrametasystem where he is its driver. A huge and complex production system has been constructed by a spiral in a way similar to our method of progressive formalization: you first create a few rough tools: set $A$; then using these tools you create a set $B$ of better tools; these better tools allow you to improve the tools of group $A$; this will be set $A'$. Then you improve $B$ using $A'$, and it goes like that on and on:

$$A \prec B \prec A' \prec B' \prec A'' \prec B'' \ldots etc.$$

(the sign $\prec$ reads: *precedes*). In the contemporary industrial system it is impossible to say what precedes what: the chicken and egg problem. But if the system is destroyed, the only way to restore it is to unwind the spiral again, starting with bare human hands. This is a characteristic feature of evolution by metasystem transitions. Mother Nature is a gigantic ultrametasystem which allowed life to develop by a spiral between nucleic acids and proteins. But we do not yet know what exactly played the role of the human hand, so we cannot create life artificially.

As in biological evolution, we can distinguish in the evolution of the production system a few MSTs of the largest scale. Take the first industrial revolution, where control was imposed on the natural sources of energy. Take the second industrial revolution: control of information and control over control itself. The evolution of computer technology can be described in terms of metasystem transitions, but I will not do it here.

**S** You know, the fact that your method of progressive formalization looks so similar to other evolving systems adds, in my mind, to its credentials. Indeed, why not use in our science and philosophy the same evolutionary principles that showed their power elsewhere?

**T** Indeed, why not. When do you think science as such started? I mean, something that is definitely not technology.

**S** At the Renaissance time, probably. Or with the ancient Greeks?

**T** I would definitely say with the Greeks. I do not separate mathematics from science, as you remember. And it was the Greeks who started mathematics for real by introducing the concept of *proof*.

Neither in Egyptian, nor in Babylonian texts do we find anything even remotely resembling mathematical proof. An equivalent of what we know as formula, but expressed partly in natural language, was known before the Greek wise men. The Egyptians, e.g., computed the area of a circle by the formula $(\frac{8}{9}2r)^2$ (which corresponds to $\pi = 3.16$). But the idea that any proposition about figures and numbers which is not completely obvious

must be proved, i.e. derived from those completely obvious propositions by a convincing logical arguing – this idea was of Greek making. Apparently, their democratic social order was responsible for it. Disputes and proofs played an important role in their life. The concept of proof already existed as a social reality; all that remained was to transfer it to the field of mathematics.

The introduction of proof is an MST within language. The formula is no longer the apex of linguistic activity. The proof is directed to the analysis and production of formulas. This is a new stage in the development of language and thought, and its emergence called forth enormous growth in the number of formulas (the law of the growth of the penultimate level). A metasystem transition always means a qualitative leap forward, an explosive development. The mathematics of the countries of the Ancient East remained almost unchanged for up to two millennia. But in just one or two centuries the Greeks created all the geometry our high school students sweat over today.

The emergence of proof was part of a larger process of movement towards *critical thinking*, by which I understand the thinking about one's own thinking. We ask ourselves: Is what I think true or false? Why do I think so? How can it be justified? Why other people might think differently? In primitive societies people accept their language, their beliefs, and their rules of social behavior as something given, like the natural phenomena. It has taken a long time, and also contacts between different cultures, for people to realize that they can think of their beliefs, analyze and change them. Philosophy, like mathematics, is the child of the metasystem transition to critical thinking.

When we look at the history of mathematics, we see, again, metasystem transitions as the milestones. Take the emergence of algebra from arithmetic. When some quantity must be found, this is an arithmetic problem, no matter whether it is formulated in everyday language or in a specialized language. And when the general method to solve a class of problems is pointed out – by example, as is done in elementary school, or even written as an equation – we still do not go beyond arithmetic. Algebra begins when the equations themselves become an object of activity, and the manipulation rules for equations and other formulas are studied. This is a metasystem transition. The formula, which defines control over arithmetic operations, becomes an object of control by the laws of algebra. Numerous new formulas are produced (the law of the growth of the penultimate level). I remember my delight when as a schoolboy I got acquainted with the basics of algebra. The arbitrary, and often vague rules which we had to use before for solving

problems were replaced by clear and fully justified algebraic transformations.

The use of abstraction, as in a step from school algebra to modern algebra, includes a metasystem transition; it is a form of modeling. The creation of formal logic and metamathematics is a large-scale MST. The proof, which was at the top level of control hierarchy in mathematics, becomes itself an object of control. The famous Gödel theorem proves that something cannot be proved.

Experimental science, as distinct from simple observation, is also a result of a metasystem transition. It is *controlled observation*, with *theory* as the control mechanism. It is like the jump from gathering to agriculture. We do not just observe nature, we ask it questions which we formulate in terms of our theories. Metasystem transitions which take place in the development of science create a multi-level hierarchy of control as shown in Fig.6. We discussed this scheme in the context of a cybernetic animal; now we deal with quite different material, sign systems, instead of nerve net, but the principle is the same. Representations in classical and quantum physics, and a possible system of metarepresentations are discussed in [3].

For human society science is what the brain is for an individual: the instrument of knowledge, i.e. of the creation of new models of reality. It is the highest level in the universal hierarchy of control, highest not in the sense that it cannot be overruled (it can; and what reason tells a human individual can also be overruled, alas, by emotions), but in its evolutionary history (the sequence of MSTs) and, therefore, its significance for future. We cannot 'overrule' the force of gravity, but the fact that things tend to fall to the ground is not a great factor defining our future; when we want we can circumvent it, as flying airplanes proves. This is the essence of control hierarchies. Teilhard stressed the cosmic importance of the phenomenon of man. I want to stress the cosmic importance of the phenomenon of science as part of the phenomenon of man.

Science is a superstructure over human brains which, though created by brain, is partly independent of it, has its own hierarchical structure and directs the work of individual human brains. Science is not simply a means to improve human condition; it is a cosmic phenomenon of tremendous importance. It is the top of the growing tree of the Universe, the leading shoot of Evolution. Immortal itself, it has as its goal the immortality for every human being.

**S** I think many people will find your apology of science exaggerated, to say the least. What about other forms of the human spiritual life, say art? You seem to write it off completely.

**T** No, I am not writing off the art, not even in its purest forms: remember what I said about esthetic education. But there is a crucial difference between science and art, which determines their long-range impact on the future. The language of science tends to be *formalized*, which means that its use can be relegated to the machine. Scientific models of the world can be separated from the human mind. It is conceivable that an intelligent Martian could understand the meaning of our mathematics and physics on the basis of our records only, as he/she/it could do in the case of a mechanical model of, say, Solar System. Because of this separation, each next level in science can treat the preceding level as objective reality. Repeated metasystem transition in the construction of model becomes possible, which results in the stairway effect and an explosive, and seemingly, unlimited development. Science walks out of the human mind, so to say.

The art, on the contrary, is inseparable from the human mind; the language of art becomes meaningless if it tells nothing to our soul. Thus there are inherent limits for the development of art, because the human body and soul remain constant – at least on the scale of cultural evolution. Science can be – and to some extent already is – *superhuman*; art is not and will never be. All modern art of our century, music, visual arts, poetry, grew out of the desire to make something really new, jump out of its own limits ...

**S** Make a metasystem transition.

**T** Yes. But the results, from my viewpoint, even though often interesting and sometime very impressive, only show once again the existence of the limits. Metasystem transition is not in the nature of art. Whatever role is played in our life by contemporary art, the classic art does not shine less, in a sharp contrast with the situation in science (who reads now the original writings of Galileo and Newton?) What is done in art – and I mean, of course, the art of all times and peoples – is done, and only so much can be added.

I want to stress once again that I am not in the least trying to diminish the role of art; I only speak about the role of new development in art and their impact on the future of mankind. While the role of *new* science, i.e. the additions to the existing science, goes up and up, the role of *new* art – let us be generous – remains constant.

**S** I am sure that the majority will disagree with you, and many will be offended.

**T** I cannot help it. It is science that shapes the future, not any other form of human activity. This is a simple fact of evolution. On this note I

announce a coffee break.

## 5 The future of the world

My picture of the future is based on my picture of the past. The present stage of Cosmic Evolution is the integration of human individuals on the planet Earth. An attempt to be more specific produces more questions than answers. How far will the integration go? There is no doubt that in the future (and perhaps not too far in the future) direct exchange of information among the nervous systems of individual people will become possible. Obviously, the integration (maybe partial) of nervous systems must be accompanied by the creation of some higher system of control over the unified nerve network. How will it be perceived subjectively? One may hope that the new level of control will result in a new, higher form of consciousness, which will come into existence on top of the consciousness of the present day individual. This new consciousness may be, in principle, immortal.

But will our descendants want physical integration? Generally what will they want? And what do *we* want, for that matter? Also, what do we *want to want*? What do we take for Good and for Evil?

These are the perpetual questions of ethics. Science, by its nature, does not give direct answers to these questions. The gap separating knowledge and will can never be fully bridged. No matter what we know, we are still free to arbitrary choose among our options. But science can provide guidance by foreseeing the results of our actions.

The evolutionary growth of the control hierarchy is a fact of the natural history which has the status of a natural law. Like every law of nature, the law of evolution does not determine uniquely and in detail how things should develop. It only sets the boundary between the possible and the impossible. No one has proved, and hardly will ever prove, that the existence of life, and specifically, highly organized life, is inevitable. We have not yet had any sign that life exists outside the Earth; as for humankind, it can destroy itself, and possibly the whole of life, if it chooses to do so. Continuing constructive evolution is a possibility but not a necessity.

No one can act against the laws of nature. Ethical teachings that run counter the general trend of evolution, i.e. set goals incompatible with it, cannot bring about a constructive contribution to evolution. This means that the deeds prompted by such goals will, in the final analysis, be erased

from the world's memory. Such is the nature of evolution: that which corresponds to its general trend, or abstract 'plan', is eternalized in the structure of the developing world; that which runs counter to it, is overcome and perishes.

It follows that if humanity sets itself some goals which are incompatible with further integration of individuals, the result will be an evolutionary dead end: further creative development and the engendering of qualitatively new forms of life will become impossible – at least with regards to our species. In such case we shall ultimately perish. In the developing world there is no repose: all that does not develop perishes.

**S** Now I am suddenly appalled by what you are saying, especially because in the coffee break I read *The Cybernetic Manifesto* by yourself and Joslyn [15]. We are given a vision of the universe as a hierarchical control system, which inexorably moves towards more and more control. At each level there is some kind of 'Will' controlling whatever lies below it. This reminds me of the economic and social system of Stalinism; the communal masses strive together to meet plans which have been set by the higher evolutionary level of hierarchical control. There are no chance events, nor even more than one possible event at any given time, since 'Will' always determines an outcome; even conflict and disagreement are ruled out be definition. This is not my idea of freedom, which has more to do with spontaneous synergistic co-operation over a field of unknown possibilities.

**T** You are appalled by a picture which you took from Orwell and Huxley, not from me. Your outcry against *The Cybernetic Manifesto* would have surprised me if I did not have a comparable experience before. But I did. A few years ago it came as a complete surprise to me that quite a number of the reviewers of my book *The Inertia of Fear and the Scientific Worldview* [13] classified my views as inconsonant with pluralistic democracy. The record was set by Major James F. Kealey. Reviewing my book in *National Defense College Quarterly*, *Joint Perspectives*, he presented me as "a dissident who, like Solzhenitsyn before him, seeks to justify a new order as totalitarian as the one he left in disgust".

The condemnation of *The Inertia of Fear* as a totalitarian book is paradoxical, because the main contents of the book is an analysis and condemnation of the Soviet totalitarianism. It also is strongly anti-Marxist: I made a point of showing how philosophical premises of Marxism translate into the terrible and miserable Soviet reality. The book was written in 1974 and smuggled out of the Soviet Union. The first half of the manuscript was confiscated by the KGB during a search of my apartment (I had to rewrite

it from scratch). I was an active member of the human rights movement in the Soviet Union. Then how did I become a 'totalitarianist'? The answer is: my book also included an outline of an evolutionary approach to social organization and ultimate human values – mostly along the same lines, as in *The Cybernetic Manifesto*. It is these ideas that some of the readers perceive as totalitarian and Stalinist. Now you joined them, even though you should have known better, because you know from our previous discussions how fundamental the idea of freedom is in this philosophy, and what is my concept of control.

Needless to say, I qualify accusations in totalitarianism as absurd, but I still have to explain why this misperception is so persistent.

The major part of the explanation is simply the superficial thinking of our critics, when conclusions are made not on the basis of what we are actually saying, but in the wake of current popular associations with the terms we use. You see our references to 'control', 'hierarchy', 'integration', and this all just sounds totalitarian to you.

When you hear 'control', you tend to imagine something like orders from the superiors, a scrutiny by a governmental agency, etc. – in any case something which should be minimized and is thus bad by its nature. But you could see from our discussion that I use this term in its most general sense. I am looking for the most general and deep aspects of the world, and one of these aspects is that the world is not completely chaotic. If the mechanistic science of the 19th century saw the notion of a deterministic natural law as the adequate expression of this idea, the cybernetic science of today puts the notion of control in its place, leaving freedom as an non-illusory and non-eliminable element. Control is not the same as compulsion. Control is only a limitation of freedom, not necessarily its elimination, and this limitation is not necessarily hurting the controlled entity. It may be life-saving, as in the case when somebody takes your hand and leads you out of a maze. Indeed, every kind of *problem solving* is a kind of control. To solve a problem usually means to pick up one true solution from a combinatorially huge number of possible false answers. You can have the full freedom to choose any answer and be very unhappy with this freedom. You wish somebody could exercise control over you by limiting your choices to only a few. Society controls children by teaching them, and adults by offering them jobs. The mother controls the behavior of the child when she kisses it.

**S** Maybe, you should have chosen another term for you general control concept?

**T** But I cannot find one. And control *is* limitation of freedom. The idea

that philosophy must treat control as evil is an outdated liberal reaction to the older idea of a rigid, mechanistic control.

Hierarchy is another victim of bad associations. If control is identified with compulsion, then hierarchy is almost always thought of as a *pecking order* – a way to dominate the weak, so we must fight against it. But a hierarchy is simply a relation of partial order in lines of control, not necessary compulsion or domination. Thus by its definition a hierarchy is the presence of some order, that is all. Often hierarchies intersect, so that one subsystem may be above another subsystem in one hierarchy, and below it in another. A colonel may command a regiment but be in complete subordination of his wife. He is also controlled by the author of the paper he is reading, and by mosquitoes when vacationing in Maine. Some of the control lines may circle (feed-back), but this does not necessarily destroy the hierarchy. Our colonel should heed to the feelings of the soldiers, but he is still in command. And when we take a feed-back loop as a whole, we would typically find that it is a part of a larger hierarchy.

Hierarchy, like control, does not exclude freedom. It is opposite to chaos, not to freedom. Organized systems are hierarchies of control, this is an observable fact. We find them everywhere.

Speaking about human hierarchies, we value a free democratic society not because it is free of control hierarchies – it has at least as many as a totalitarian society, but because the character of control is different. The industrial hierarchy, for example, is controlled by free market, state regulations, bank financing etc., but not through direct administrative orders, as in the former Soviet Union. In fact, the Western economical and political system is much more sophisticated and includes much more intersecting hierarchies than the primitive Soviet system. In a word, it is more 'cybernetic' than the 'mechanical' Soviet system. This demonstrates that a society can be both more free and more 'hierarchical' – in the above sense, without confusing the presence of hierarchies with the control through compulsion.

**S** Yes, I understand your point, but when one reads a short document like the *Manifesto*, one gets scared. Perhaps you should make more emphasis on the place of freedom in integration.

**T** I was going to do this anyhow. In fact, we did this in the Manifesto, if you read it carefully!

Freedom and integration. We need both. Integration is an evolutionary necessity. Let me repeat. I believe that if humanity sets itself goals which are incompatible with integration the result will be an evolutionary dead end: further creative development will become impossible. Then we shall

not survive, because in the evolving Universe there is no standstill: all that does not develop perishes.

On the other hand, freedom is precious for the human being; it is the essence of life. The creative freedom of individuals is the fundamental engine of evolution in the era of Reason. If it is suppressed by integration, as in totalitarianism, we shall find ourselves again in an evolutionary dead end. This contradiction is real, but not unsolvable. After all, the same contradiction has been successfully solved on other levels of organization in the process of evolution. When cells integrate into multicellular organisms, they continue to perform their biological functions–metabolism and fission. The new quality, the life of the organism, does not appear despite the biological functions of the individual cells but because of them and through them. The creative act of free will is the 'biological function' of the human being. In an integrated super-being it must be preserved as an inviolable foundation, and the new qualities must appear through it and because of it. Thus the fundamental challenge that the humanity faces is to achieve an organic synthesis of integration and freedom.

**S** So what is then your ethical principle which you derive from your cybernetic philosophy?

**T** The Supreme Good is constructive evolution spearheaded by science. Constructive evolution includes further integration of human society combined with preservation and enhancement of creative freedom of individuals. In perspective, the achievement of immortality of human, or human-like, beings. Immortality needs to be discussed separately.

I very well remember that moment in my childhood when I clearly realized for the first time that sooner or later I will die – inevitably. It is not a matter of intellectual understanding, but rather the work of imagination. Lev Tolstoy, and some other authors describe it. I believe most people had this experience at some time. What about you?

**S** I think I know what you are speaking about.

**T** This is a terrible feeling, worse than pain. It comes as a shock. You feel that you are cornered, and there is no way out. Your imagination jumps over the years you have still to live through, and you find yourself on the brink of disappearance, complete annihilation. You realize that you are, essentially, on the death row. Different individuals react to this situation with different degree of pain. Some simply try to forget about it, and succeed, to some degree. Others try forget but cannot. Life seems to have no point, because all roads lead to annihilation; one is haunted by the feeling that whatever he is doing is in vain.

The realization of one's own mortal nature is one of the most fundamental distinctions between a human being and an animal. *The will for immortality*, a rebellion against death, is found at the source of religions, philosophies, and civilizations. People look for a way to transcend the limit put on our lives by nature. They look for a concept which would reconcile the impulse to live on, which is inherent to every healthy creature, with the inevitability of death. Some concept of immortality becomes necessary for keeping life meaningful, and I can count four.

The immortality as understood in the classical religions I designate as *metaphysical*. It is referred to as immortality of soul, life after death, etc. Metempsychosis, the lore of migration of souls, is also a variation on this theme. The basic feature of metaphysical immortality is that it is limited to the conceptual sphere. No physical reality takes part in forming this concept. In fact, the concept defies physical reality and proclaims – without a hint of proof – the reality of a different kind. Traditional religious teachings begin from an unconditional belief in the immortality of the soul. In this case the protest against death is used as a force which causes a person to accept this teaching; after all, from the very beginning it promises immortality. Buy under the influence of the critical scientific method, the notions of immortality of the soul and life beyond the grave, which were once very concrete and appealing, are becoming increasingly abstract and pale; old religious systems are slowly but surely losing their influence.

I will call *creative immortality* the idea, in one form or another, that a mortal human being contributes something to the ongoing universal and eternal process, which can be called Evolution, or History, both with capital letters. I call it Evolution, because contemporary science tells us that human history is but a small part of the universal cosmic process.

The concept of Evolution provides a link between knowledge and will which can serve as a basis for distinguishing between good and evil. The contribution to the Evolution made by an individual can be of critical importance. It can also be everlasting. The contributions made by Aristotle or Newton are written down into the history of mankind and will stay there forever, even though there are only very few people who read Aristotle or Newton now. This is because each next stage of evolution is dependent on the preceding stages. The acts contributing to evolution create structures which will outlive the actors and determine the structures that follow. In this way, they are eternal. Creative immortality may also be called *evolutionary* immortality; it is the immortality of deeds. The deeds of mortal men may be immortal.

The concept of a simple *biological immortality* is as easy to understand as it is hard to implement. I am speaking of the infinite continuation of individual life in the same form as we know it, i.e. based on the same biochemical processes in our bodies that make us living now. There is a mechanism of ageing and death which is built into our bodies by nature. If we could somehow switch off this mechanism, we could, in principle, live indefinitely long. Our life is based on a metabolism; the body has a capacity of self-renewing. The process of life could be, in principle, unlimited in time.

However, there are some qualification to this concept. First, there are still chances of an accidental death, which become the more serious the longer we live. Second, contemporary biology does not yet answer definitely whether the infinite (or very long; say, hundreds of millions of years) life is feasible. It is possible that the mechanism of ageing is built-in on such a deep level, that you cannot switch it off without radically altering the whole machinery of bodily life.

This brings me to the last concept: *cybernetic immortality*. For the time being we find it only in science fiction. The idea of cybernetic immortality is that the human being is, in the last analysis, a certain form of organization of matter. This form is distinguished by a very sophisticated organization, which includes a high multilevel hierarchy of control. What we call our soul, or our consciousness, is associated with the highest level of this control hierarchy. This organization, which we associate with our 'I' can survive a partial — perhaps, even a complete — change of the material from which it is built. Moreover, it can infinitely evolutionize, become even more sophisticated, and explore new, yet not thought of, possibilities. Even if the decay of biological bodies is inevitable, we can look for some ways of information exchange between bodies and brains which will preserve in some form the essence of self-consciousness, our personal histories, our creative abilities and, at the same time, make us part of a larger unity embracing, possibly, a huge number of human individuals.

This aspect of cybernetic immortality, integration of individuals, seems to be its inevitable component. The exchange of information between brains through the channels of sense organs is, cybernetically, extremely imperfect. More direct forms of exchange signify a much higher degree of integration than we can speak of at present. Direct exchange will give tremendous advantages in terms of intellectual strength. By the evolutionary law of survival, human conglomerates exercising such exchanges must proliferate and seize the top level of control in the world. Also, from the view of personal immortality some kind of integration is inevitable: If your soul is to be stored

somehow in a form similar to computer software, you will want the storage to be attended, properly updated, and preserved with some redundancy to fight accidents.

**S** How do you envision the process of ultimate integration?

**T** OK, let us speculate. Judging from the history of evolution, it is unlikely that the whole humanity will unite into a single super-human being. Even though human beings have seized control in the biosphere, they make up only a tiny part of the whole biomass. The major part of it is still constituted by unicellular and primitive multicellular organisms, such as plankton. Realization of cybernetic immortality will certainly require some sacrifices — a vehement drive to develop science, to begin with. It is far from being evident that all people and all communities will wish to integrate into immortal super-beings. It is probably as certain that this *will not* happen as it is certain that some individuals, and then communities, *will* set this as the supreme goal. The will for immortality, as every human feature, varies widely in human populations. Since integration we speak about can only be free, this means that only a part of mankind will be involved in integration. The rest will continue to exist in the form which in the *Manifesto* we called 'human plankton'.

**S** Which, again, will be resented by many. I see here the division of people into a higher race, which will be integrated and will become the race of masters, and the lower race which will not be integrated and will exist as the race of slaves, 'human plankton'.

**T** Well, this is a result of my respect for your freedom. If you do not want to integrate, do not, but be prepared to accept the consequences. There is an alternative: to compel everybody to integrate, but you certainly would not accept this either.

**S** Certainly not. But I will prefer another alternative: not to integrate at all. The life of the mortal human individual, as I know it, suits me very well, and will suit by children and children of my children. And this, I am sure, is the feeling of the vast majority.

**T** Oh, yes. This line of thought is familiar. Now *you* want to limit *my* freedom to integrate. Fundamentalists of various kinds agree on one point: to curb science so that it does not ...

**S** I am not a fundamentalist.

**T** In some sense you are; only you have different scriptures. Well, call it conservative. Since you tend to block the road of evolution as seen through the eyes of science ...

**S** But this is not the eyes of science, this is seeing it through *your* eyes!

**T** Quite true. But I am trying to make next step on the basis of what has been firmly established and agreed upon. If you disagree, give your picture of evolution.

**S** Why should I? Evolution is a result of the actions of billions of people, as well as natural causes. Let it proceed as it does.

**T** In other words, do not interfere with Divine Providence. I told you you are a fundamentalist. You are washing your hands. And your justification is that the future does not depend on what any single individual is doing. I call this unethical. Because your justification is false. You imply that only global factors matter, only averages over big numbers of individuals make a difference. But the use of the law of big numbers would be justified only if the actions of those individuals were independent. They are not, of course. Society is a tightly bound system, and we know that in such systems trajectories in the configuration space diverge: small variations in the present may lead to huge differences in the future.

**S** Speaking of justification, you cannot call anything unethical without really justifying what you believe to be ethical.

**T** This is quite true again. But the question is: what kind of justification can be found? There is a great diversity in human society. You seem to have a gut feeling against cybernetic immortality and integration, and there are many people who feel similar. But there are also many people who would have a gut feeling *for* cybernetic immortality. I can point at Elan Moritz's paper in this issue (see [7]). Moritz discusses these ideas in the context of the theory of *memes*. He uses the term *a Postorganic Immortal Persona*, which speaks for itself.

So let me repeat what I said before: the gap between knowledge and will cannot be fully bridged. You are free to choose any ethics, any idea of the Supreme Values. The ethics based on the theory of evolution can only tell you how you can hope to achieve immortality. It appeals to your *will for immortality*. If you do not want it, you can do whatever you please.

**S** You mean 'creative' immortality.

**T** Cybernetic immortality, too.

**S** But this is only for the people in a distant future.

**T** Not necessarily. You know that there are organizations which accept human bodies for keeping them at the liquid nitrogen temperature until the time when it becomes possible, because of the development of science, to revive and cure them. If one believes that cybernetic immortality will ultimately become a reality, one can, in principle, keep one's body, or only the brain, in that way, and ultimately become immortal.

**S** But this is absurd.

**T** Why? As long as the information identified with what we call one's soul is not lost, it should be possible to launch a form of life which will be a continuation of the person's former life. This is, of course, a question of faith, but to believe in it is not, from my view point, absurdly. There is short-range and long-range memory. Even if short-range memory is lost at clinical death, the long-range memory may, probably, be preserved at the liquid nitrogen temperature for many hundreds of years.

**S** The crystals of ice which are formed in freezing damage the cells.

**T** I know. But can you prove that nobody will ever discover a way to solve this problem? Certainly, not. This is why I say that this is a question of faith, but this faith is not absurd.

**S** You, apparently, believe in cybernetic immortality. Will you instruct to keep your dead body frozen?

**T** I am afraid it will be too expensive for me. Immortality costs money. But I can imagine that in some time some people and organizations will emerge which will work to make cybernetic immortality real, and they will preserve their bodies for reincarnation.

**S** This would cause most severe social problems... But I, really, do not take this seriously. This is just a fantasy.

**T** It may seem like a fantasy right now, but recall how many things were just a fantasy before becoming one more miracle of science. When I was a boy, I read a lot of science fiction about space flights, and I would have never believed that man would be on the moon before I am forty. I think you underestimate the impact that the idea of real immortality can have on the human race if it at some moment it becomes taken seriously. From the emergence of human beings our history has been a history of surviving. Now the time is close, hopefully, when due to science and technology the whole human world is integrated, peaceful and thriving in the bounds set by biology. Then what? What great purpose will the human being set for itself? To overstep those bounds, of course! Immortality is the only *natural* supreme goal – in the sense that it is not invented and does not rest on blind faith. The will for immortality is a continuation in a thinking creature of the animal will to survive.

**S** You refer all the time to evolution and its *natural* course, but what you defend is highly *unnatural*, it threatens to destroy human life as we know it with all the beauty of human body and mind created by nature.

**T** Not at all. There is a general law of evolution: ontogenesis, the history of an individual development, roughly repeats, or recapitulates, phylogen-

esis, the history of the species. Evolution tends to build on the existing foundation, making only those alterations that are necessary. Applying this law to the post-biological development we visualize a future when human beings are born and grow up in the same way as now, more or less. It is only later, probably when ageing becomes a problem, that they establish a direct cybernetic contact between their individual nervous systems and the super-brain of the human super-being. As time goes on and the biological body disintegrates, the individual mind – or soul – becomes only a part of the Supermind – or the Oversoul. Cybernetic immortality come not *instead* of our normal life, but *in addition* to it, instead of death.

I will go further and draw an analogy between the human person and a gene. In natural selection the source of change is the mutation of the gene; nature creates by experimenting on genes and seeing what kind of a body they produce. Those bodies are selected that better preserve the genes. The journal of progress is kept in genes, thus genes are immortal. As for our bodies, and therefore, minds, they are expendable; nature does not care about them. Biologically we are mortal. In the evolution of the human superbeing it is the creative core of the human individual that is the engine of evolution. Therefore evolution must make it immortal, as it made genes immortal at the preceding level of development. Brain was an object of experimentation in the biological era of evolution. At the present stage of evolution, it is the source of creativity, not an object of experimentation. Its loss in death is unjustifiable; it is an evolutionary absurdity. The immortality of human beings is on the agenda of Cosmic Evolution.

**S** Objection. You speak of nature as if it were a thinking entity which calculates how it should proceed. In fact, evolution is a result of the interplay of many blind actions.

**T** Objection sustained. You should understand this argument metaphorically. But it can be easily translated into a more strict language; it is quite convincing to me.

**S** The chain of deaths and births makes it possible for new brains to adjust to new situations. If it is stopped, you will face millions of old people who are unable to get beyond the notions of their youth.

**T** First, I did not say that the *births* will stop. I assume that the Universe will accommodate for more and more human (superhuman) souls in the foreseeable future. Second, the individual soul will not stay unchanged and petrified, because it will not me implemented in the current biological material. We can, again, understand this situation by analogy with the level of genes.

Genes are controllers of biological evolution and they are immortal, as they should be. They do not stay unchanged, however, but undergo mutations, so that human chromosomes are a far cry from the chromosomes of viruses. Cybernetically immortal human persons may mutate and evolve in interaction with other members of the super-being, while possibly reproducing themselves in different materials. Those human persons who will evolve from us may be as different from us as we are different from viruses. But the defining principle of the human person will probably stay fixed, as did the defining principle of the gene. The new will be a superstructure on this basis, as is typical for evolution.

**S** By the way, genes are not immortal. They decay together with the body. And in the case of a bad mutation they become extinct.

**T** We speak of organization or form of these things, not their implementation in concrete markable objects. Atoms are identical. When a gene reproduces itself, we say that this is *the same* gene. As for mutations, the constancy of an object is always relative. It remains 'the same' as long as only small changes take place. An object is its history. (Remember my ontology?) This applies also to human soul. It is the sudden disintegration of the soul in death that I find unjustifiable.

**S** Do you consider the possibility that man could be the last link of biological evolution and the stepping stone of "mineral" evolution through our ever more sophisticated artificial creatures which may in the end eliminate human beings as a hangover of the distant past? This is one of the most pervasive themes of science fiction.

**T** In some sense "mineral" evolution has already started: look at millions of people who carry pacemakers. But it is extremely unlikely that humans will create independent artificial creatures which will exterminate their creators. That would be against the essence of evolution, which is building new developments on the basis of the already existing achievements. The future belongs to man-machine combinations which will be more human beings than machines (by a machine I mean here any artificial component), because they will be "improvements" of human beings. Nothing human will be lost: there is no reason for it. Evolution is an on-going search for better and better solutions of the problem of stability. In the last analysis it is what in computer science is known as exhaustive brute-force search, whether it occurs naturally, or is set up by scientists and engineers. Evolution of life has been going on for billions of years on the scale of the Earth. To throw away its achievements and start from scratch as independent "mineral" beings which could ultimately overpower man-machine combinations? It seems

impossible. At least, this is much less probable than for us to perish from our distant cousins, some especially malicious bacteria or viruses.

**S** Oh, I wanted to note several times already that you cannot speak of human immortality anyway, because the whole Universe may come to an end.

**T** That is all right with me. I will be quite satisfied with a life span of a few billion years. What I am after is a cosmic role for the human race or, rather, its integrated part. No such role is possible without integration. Can we imagine 'human plankton' crowded in space vehicles in order to reach a distant star in ten, twenty or fifty generations? The units that take decisions must be rewarded for those decisions. Only integrated immortal creatures can conquer the outer space.

**S** Meanwhile?

**T** Meanwhile I believe that now more than at any time we need an integrated world view, a comprehensive philosophy based in the modern science, and specifically, cybernetics. It is necessary for future development of science itself. And most important, it should give answers to the fundamental to every human being questions about the meaning and the goals of life. The problem of ultimate values is the central problem of our present society. What should we live for after our basic needs are satisfied by the modern production system? What should we see as Good and what as Evil? Where are the ultimate criteria for judging social organization?

Historically, great civilizations are inseparable from great religions which gave answers to these questions. The decline of traditional religions appealing to metaphysical immortality threatens to degrade modern society. In fact, it *is* degrading. Cybernetic immortality can take the place of metaphysical immortality to provide the ultimate goals and values for the emerging global civilization. We shall badly need it, I believe, in the immediate future. As for a more distant future, it will be defined, if I can predict, by the conflict between integrationists and the rest of society. Such are the ways of evolution. Integrationists will be denounced from both the right, and the left. The conservatives will cry murder. Those now called liberals will cry elitism and totalitarianism. God only knows what will come out of it, but I believe in evolution. Those who think that the history is about to end are mistaken. The real history of mankind is only beginning.

**S** This sounds as a conclusion, and it should. Thank you. It was interesting, even though I disagree on some important points. I also must note that your exposition has often been sketchy, and on many occasions it was not exactly clear what you meant.

**T** This is true. But I wanted to give a review of the system as a whole, hence the sketchy character of it. You also should be aware that I do not speak of a completed work. This is only an outline of the most salient points as I see them at present. Further collective work is needed to actually bring our philosophy to an acceptable form. And by the nature of a philosophy that puts evolution above all, our work must undergo continuous development, and thus never be fully completed. Thank you, and let us hope that we will have more occasions to meet and continue our discussions.

**Acknowledgment**

It is my pleasure to acknowledge that from the inception of the Principia Cybernetica Project in 1990 constant discussions with Francis Heylighen and Cliff Joslyn have been very important for sorting out, clarifying and developing the ideas presented in this work.

## References

[1] Campbell, Donald T., Evolutionary Epistemology, in: P.A.Schilpp (Ed) *The Philosophy of Karl R.Popper* LaSalle, Ill, Open Court Publishers, 1974, pp.413-463.

[2] Metasystem Transition Schemes in Computer Science and mathematics, in this issue.

[3] Heylighen, F., *Representation and Change*, Communication and Cognition, 1990.

[4] (Meta)Systems as Constrains on Variation: emergence and evolution, in this issue.

[5] Heylighen,F., Joslyn, C., and Turchin V. A Short Introduction to the Principia Cybernetica Project, *Journal of Ideas*, Vol.2, No 1, pp.26-29, 1991.

[6] Joslyn, C., in this issue.

[7] Moritz, E. Metasystem Transitions, Memes, and Cybernetic Immortality, in this issue.

[8] Pattee, H. (1982) "Cell Psychology: An Evol. View of the Symbol-Matter Problem", *Cognition and Brain Theory*, v. **5**, pp. 325-341, 1982.

[9] Popper, K. *Objective Knowledge: An Evolutionary Approach*, Oxford University Press, 1972,1979.

[10] Powers, W.T. *Living Control Systems*, The Control Systems Group, Inc., Gravel Switch, Kentucky, 1989.

[11] Teilhard de Chardin, *The Phenomenon of Man*, transl. by B.Wall. Harper and Row Torchbook ed., 1965; p.36.

[12] Turchin, V.F. *The Phenomenon of Science: a cybernetic approach to human evolution*, Columbia University Press, 1977.

[13] Turchin, V.F. *The Inertial of Fear and the Scientific World view*, Columbia University Press, 1982.

[14] Turchin, V.F. A constructive interpretation of the full set theory. *Journ. of Symbolic Logic*, 52, pp.172-201, 1987.

[15] Turchin, V.F. and Joslyn, C., The Cybernetic Manifesto, *Kybernetes*, Vol.19, Nos.2 (pp.63-64) and 3 (pp. 52-55), 1990.

[16] Turchin, V.F., On cybernetic epistemology, *Systems Research*, 10, No.1, pp.3-28, 1993.

[17] Turchin, V.F., The cybernetic ontology of action, *Kybernetes* 22, No. 2, pp.10-30, 1993.