The Mental Life of Artifacts. Explications, Questions, Arguments*

Wolfgang Buschlinger[a], Gerhard Vollmer[a], Henrik Walter[b]

[a] Seminar für Philosophie, Technische Universität Braunschweig, Geysostraße 7, D-38106 Braunschweig, Germany
[b] Abteilung für Psychiatrie III, Universität Ulm, Leimgrubenweg 12–14, D-89070 Ulm, Germany

Z. Naturforsch. **53c**, 455–479 (1998); received May 5, 1998

Algorithms, Artificial, Explanation of Mental Properties, Reduction, Realization vs. Simulation

Our working assumption reads: Every mental property exhibited by a natural system can be exhibited by an artificial system, at least in principle. Is this true? The paper is in two parts. The first part is explicative: What do we mean by 'natural' or by 'artificial'? When are we ready to ascribe thinking to a system? We show that sometimes behavior is enough, sometimes underlying mechanisms are decisive. We study the difference between simulation and realization of a property and the kinds of explanation used in the philosophy of mind. The second part is argumentative. We collect and discuss positive arguments for and critical arguments against the working hypothesis. No decisive counterargument is found.

Studying natural and artificial systems, we sooner or later run into the question: Which achievements of natural systems can artificial systems ever reach? This question not only challenges scientific *research*, but also leads to controversial *debates* about the limitations of artificial systems. Where do these controversies come from? There are, at least, two distinct sources of misunderstanding: pitfalls of language, and disagreement on (the quality of) arguments.

The present paper centers on these two sources of controversies. While chapters A and B investigate explicative problems, chapters C and D study arguments. In A, we strive for explications: What is natural? What is artificial? What's the working assumption of this paper? (A1) How do we solve terminological problems? Is there a standard way to answer "What-is?" questions? (A2: No!) In B, we focus on the difference between simulation and realization (B1), and on the types of research characteristic of our subject (B2). Given our working assumption, we collect and discuss positive arguments in C and counterarguments in D, laying emphasis on the counterarguments.

A The Quest for Meaning

A 1 How natural may artificial systems be?

a) 'Artificial organisms' – pure nonsense?

Reading the title of this conference, you will spontaneously come up with an objection: Whatever might be meant by 'natural' and 'artificial', is it not evident that organisms are natural? And vice versa: Is it not evident that what is natural cannot be artificial? Thus, while the expression 'natural organism' is the same thing twice over, hence pleonastic or *tautological*, 'artificial organism' is *self-contradictory*, inconsistent, pure nonsense. But if artificial organisms don't exist, there is no point in talking about their brains. What then is the conference all about? A similar problem arises with the recent concept of 'artificial life', built by analogy with 'artificial intelligence'. But this is not our problem here.

A similar objection might read: Mentality is some kind of inner aspect. As far as anybody knows, mentality is restricted to organisms, to higher animals, maybe to primates, or even to human beings. No artifact so far, be it machines, or computers, or robots, has ever shown mental properties. Thus, it is often claimed that a mental life

* This communication is a contribution to the workshop on "Natural Organisms, Artificial Organisms, and Their Brains" at the Zentrum für interdisziplinäre Forschung (ZiF) in Bielefeld (Germany) on March 8–12, 1998.

Reprint requests to Dr. W. Buschlinger.
Fax: (0531) 391 81 61.
E-mail: w.buschlinger@tu-bs.de

0939–5075/98/0700–0455 $ 06.00

of artifacts does simply not exist and certainly never will. So what may this paper be about?

Well, attracting your attention might be one goal of a provocative title. Setting you thinking might be another. Hinting at an interesting problem a third. Eliciting a discussion on that problem a fourth. And promising a solution to that problem, at least making some progress, a fifth. These are the aims of this introductory paper.

Having got your attention, we try to make you realize that there are, indeed, interesting problems worth thought and discussion. We shall not, however, solve all these problems. But we shall try to show that there is, indeed, some sense in concepts like 'artificial organism' and 'artificial mentality'.

b) What's natural? And what's artificial?

Answering these questions might seem a simple task, but indeed it's not. In order not to get lost in an endless chain of definitions, let's presuppose some intuitive notion of 'system'. In a first attempt we would call *natural* all systems whose existence, development and influence is independent of human intervention. *Artificial* then, or *man-made*, is everything whose existence *or* development *or* influence on the world depends on human intervention.

Although this explication is adequate for inanimate systems, it is not for living systems. Consider children: Although their existence is obviously bound to human intervening, as is their growing and maturing, they are not artificial at all, but rather quite natural. The same is true for cattle breeding: A calf might admittedly come into existence by artificial insemination; the result, however, is not an artificial, but rather a natural system, just as children are. The calf case exhibits even better what makes a system natural: simply being alive. Obviously, organisms, even if coming into existence artificially, are natural systems. When using the term 'living' we do not just mean systems with metabolism and reproduction, but systems made from flesh and blood, or cells at least, or protoplasm. Thus, it seems that both the organismic stuff *as well as* its functioning are decisive.

A telling illustration of this intuition can be found in what Putnam calls "Ziff's argument" against the possible consciousness of a given high-end robot: No robot will ever be conscious. Why? Obviously, a conscious system has to be alive. And no robot is! Just open it and all you will find are chips and wires, and nobody would say that the robot is alive. Putnam reports Ziff's supposed position: "Even if a thing looks like a flower, grows in my garden like a flower, etc., if I find upon taking it apart that it consists of gears and wheels and miniaturized furnaces and vacuum tubes and so on, I say 'what a clever mechanism', not 'what an unusual plant'. It is *structure*, not *behavior* that determines whether or not something is alive" (Putnam, 1975). According to this view, being natural is a necessary condition for being alive. And what is artificial, cannot be natural. Thus, to speak of artificial life must be semantical nonsense.

This view is obviously based on a semantical argument. However, to argue semantically is often dissatisfying. Why should and how could our linguistic conventions determine which properties artificial systems might have and which they might not have? In fact, for many concepts this is *not* the case. Consider a famous example: the concept of 'number'. For the Pythagoreens, the number one was not a number, but rather the *constituent*, the generator, of all numbers, just as a brick is a constituent of a wall, but never a wall itself. Nor were there negative numbers, or a number zero. What on earth should they ever count or number? For quite some time, some other numbers acknowledged today went as *unreal* or *imaginary*. Obviously, the former use of the term 'number' did not prejudge once and for all what was ever to go as number. Nor did its etymology, nor any authority.

What is true for numbers, could also be true for consciousness. To exclude artificial systems conceptually from being conscious just because they do not have the material make-up we are used to reminds of excluding the negatives from being numbers for not denoting anything numerable.

What is more, even if we were to accept Ziff's argument and to agree that *robots* cannot be conscious because they are nothing else but clever mechanisms in Ziff's sense, we need not abandon the hope for artificial consciousness in general. "Well," we might argue, "let's agree that *robots* will never have conciousness because they are not alive in your sense. But could there not be other artificial systems with consciousness? What about *androids*? They are, though artificial, both alive

and conscious." Tellingly enough, the system *Data* in "Star Trek" is not a robot, but an android.

Thus, there is no necessary contradiction between being alive and being artificial. Therefore, Ziff's argument does not count against an artificial system as being conscious, but only against a robot being conscious, and this, in turn, only due to a rather restrictive terminology – which might be changed.

c) Our working assumption

Being conscious is, of course, a high-brow property. It is imaginable that artificial systems, for example computers, cannot be conscious, but can still have *other* mental properties. There are many candidates: sensations, feelings (aesthetical, moral, or whatever), emotions, curiosity, perception, knowing, imagination, willing, intention, thinking, reasoning, humour, and other mental properties. Though they differ in kind, scope, distinctiveness, function, all these properties have one trait in common: There is some kind of inner aspect to them. We might try to list them completely, to systematize them, to build a hierarchy, to look for the most elementary one. None of these is our task here. Rather we ask: Whichever mental property there is, may an artificial system have it? As a provisional answer, we formulate our working assumption.

The working assumption reads: *Every mental property exhibited by a natural system can be exhibited by an artificial system, at least in principle.*

Or, to put it slightly different: *Artificial systems can have or realize the mental properties of natural systems.*

Obviously, this is a strong claim. We could be more modest. We could refer to one mental property only or to just a few of them. But which one(s)? High-brow or elementary? For now, we shall stick to the more ambitious claim. Of course, we cannot prove it. Nobody can. But can you refute it? The point in formulating such a strong claim as a working assumption is not to be right, but to provoke arguments and thereby progress to better assumptions.

Arguments are delayed to chapters C and D. For now, we want to understand better what our working assumption amounts to. It does *not* claim that any artifact *is* itself natural. Nor does it claim that one single artifact should have *all* mental properties natural systems have. In that respect, it is not the most ambitious claim thinkable.

It is ambitious, however, with respect to the manner the artifact shows those mental properties. This manner is not restricted to *simulation*. We don't say that the supposed artifact just *feigns* to have mental properties, that it might give the (possibly false) *impression*, or that it might successfully *deceive* us. We rather say that it may *have* pain, or ideas, or intentions. Ascribing mental properties to an artifact is not meant to be erroneous, but to be correct, not meant as a loose way of speaking, but as a legitimate one, not in a vague metaphorical sense, but in some proper sense which, of course must be specified.

This, then, is our next task: thinking about our terminology. How are meanings established? How are they changed? Is there a *standard* way we could or should rely on? What do we mean by having or exhibiting a property? What is the difference between having a property and simulating it? When do two systems have the same property? When are two systems equal or equivalent? Is behavior or performance enough, or do we have to care for mechanisms? How deep must we dig in order to ascribe a property correctly?

A 2 From flying to seeing to thinking

a) What do you mean by it?

Once we agree on something there is no need for further discussion. Nor is there need for philosophy. But we disagree on so many things, and in principle we may disagree on everything, even on the question whether we disagree. In cases of disagreement, we ask further questions: *What do you mean* by 'democracy'? *How do you know* that the earth moves around the sun? *Why should I* not tell lies? In such cases philosophy may be helpful. Philosophers ask questions nobody else might ask at that time. They look for problems, for questions, for possible answers. But first of all, for arguments. And counterarguments. That's why they will have such things in stock if they are needed. Doing philosophy means thinking in advance (Denken auf Vorrat). Let's see whether we can take profit from their stocks.

The question "Can machines think?" has been answered 'yes' and 'no'. There is no agreement on

this question. Thus, we normally prefer to ask back: What do you mean by 'think'? We may even ask: What do you mean by 'machine'? Or by 'can'? But the first question is the hardest one. And then it's not enough to say: Well, by 'think' I mean what everybody means by it. Nor is it enough to say: I mean exactly what Plato meant. We want an *explication*. (Mind: An explication is not an explanation, but some kind of definition, a pinning down and sharpening of meaning.) It need not be ultimate, but at least workable.

What do we expect from a workable explication? It should be *understandable*; it should go as a *formally correct* definition; it should be as *precise* as necessary for our actual goals; it should cover the *usual* meaning, everyday or scientific; but it may also *deviate* from this for the benefit of other virtues. Thus, there is an interplay between a *descriptive* element (how is the word actually used?) and a *conventional* element (how *should* we use it?), between stating (fest*stellen*) and stipulating (vereinbaren, fest*legen*). This interplay is the battleground of age-old and never-ending discussions between essentialism and conventionalism, between foundational and pragmatic approaches.

Some concepts are easily explicated, others are not. Let's review some of the more difficult concepts, arranging them according to the trouble their explication will make us.

b) When do we say that a system is flying?

Leaves fly with the wind. Cannon-balls fly in parabolas, planets in ellipses. There are flying fishes, flying frogs, flying squirrels. There are birds, bats and flying insects, especially flies. But there are also gliders and kites, flying boats and – why not? – flying saucers. You may fly a balloon, a Zeppelin, an aircraft, a jet, a helicopter, even a rocket. You may fly to the moon. In sum, there are both flying organisms and flying machines. Since flying artifacts are, historically speaking, quite recent, the idea of flying, the very *concept* of flying must have been transferred from animals to machines.

But does a ballon *really* fly? If you have the principle of aerodynamic lift in mind, you would surely give a negative answer to the question: No, ballons do not really fly, but planes do. However, is this answer doubtlessly correct? Does a plane *really* fly? A fundamentalist might be tempted to deny that. The *essence* of flying – she might argue – lies in moving wings upward and downward. Planes do have wings, but they don't move them; therefore they do not fly. Daedalus tried to fly, but didn't succeed. The tailor from Ulm tried, but fell into the Danube. Lilienthal tried and semisucceeded, but finally failed and died. Men just cannot fly. Machines could in principle, if they were artificial birds with wings, muscles, feathers, pinions. They would have to imitate nature close enough. But what is close enough?

Apparently, everyday and even technical language has taken another course. We don't hesitate to ascribe flying to aircraft. The principles by which flying is realized are numerous. For ballistic weapons it's inertia and gravity, for balloons it's buoyancy, for planes it's airscrew propulsion and aerodynamic lift, for rockets it's repulsion. For most flying systems, especially for birds and planes, air is necessary; they are literally *airborne.* Not so for rockets: They don't need air or gravity. Do they, then, *really* fly? Obviously, our everyday terminology doesn't care for the principles; it cares for results, for performance, for achievements. In this case, our terminology is rather generous: Flying is ascribed to many things, even to artifacts, even if their flying is not based on the same principles. And it would be insane to try to "purify" our language and to restrict the term 'flying' to birds, or to animals, or to airborne motion.

c) Handling Rubik's cube

Rubik's cube consists of 26 small colored cubes outside and one sophisticated supporting cube inside such that any slice of nine cubes may be turned against the others. The task is to rearrange the cubes such that every side is uniformly colored. To do it by trial and error is virtually impossible. A mathematician could use group theory in order to fix the cube. There are, however, recipes such that neither trial and error nor an understanding of group theory are needed although the recipes may result from a group theoretical approach. These recipes are so simple that provided with them even a six-year old child may do it in one or two minutes. Having spent hours, days or weeks without success, and then realizing that it can be done so quickly, you come to believe in

miracles. In this case, we don't care for background knowledge, for group theory, for mechanisms, but just for *performance.*

d) When is a system calculating?

Men do calculate. By calculating we mean handling numbers: adding, subtracting, multiplying, dividing. We learn it at school. Some people use their fingers. Do fingers calculate on their own? No. Does a slide rule multiply? No. Nor does an abacus or a logarithmic table on its own. It's rather *people* who calculate *using* fingers or slide rules or abacuses or tables, but fingers, slide rules, abacuses, tables don't. The latter are tools (Werkzeuge), thinking tools (Denkzeuge), as we may call them, but not calculators on their own. Now, if someone solves multiplication problems by looking up results in a multiplication table, does she really calculate? No. Again, if she has learned the table by heart and uses her memory, does she calculate? Another no. In this case, performance is not enough: We care for the way the performance is achieved.

This does not preclude the use of tools, be they material or mental, be they machines or algorithms. Algorithms are problem-solving procedures. They serve for reliably reaching results. They are, so to speak, abstract short-cuts. We like short-cuts. Thus, we don't mind machines using short-cuts. At least, we shouldn't. We shouldn't even mind machines using methods precluded to organisms. There are, for example, no airscrews in nature (or nearly none) because there are no wheels in organisms (or nearly none). Even so we accept planes, being propelled by airscrews, as flying. Likewise, we should not deny the ability to calculate to a system just because the way it gets its results is different from ours.

May apes calculate? So far as we know, they don't and can't. They count up to eight, as ravens and jackdaws do; but counting is not calculating. No doubt they could be trained to memorize a small multiplication table, up to four by four, let's say. (We don't know whether this experiment has been done.) But still, they would not be calculating. Performance is not enough.

How about a pocket calculator? Does it calculate? In a sense, it would be absurd to deny that. A calculator that doesn't calculate? How strange! But does it *really* calculate?, we might be tempted to ask. May a Turing machine, may the Analytical Engine, Babbage's mechanical computer, ingeniously designed, but never finished, calculate? Does a computer, essentially being a universal Turing machine, calculate? When humans calculate, they *know* that they are calculating. They know what numbers are, and they know what sums, differences, products, quotients are. Is this knowledge essential? If yes, then a pocket calculator would not calculate. Nor would a computer. If no, they would.

Obviously, it's up to us, whether we insist on this additional knowledge element. Again, everyday and technical language have since long decided the question: Babbage's engine was designed to calculate, pocket calculators do in fact calculate (but not much else), computers calculate (and they do much more). According to our language use, the additional element of consciousness is *not* necessary.

e) When does a system prove something?

Men can prove theorems. At least mathematicians can. As far as we know, it was up to the Greeks to discover that so many theorems of geometry may be reduced to a few elementary axioms (and logic). The quest for proof has swept all over mathematics, if not all over science. In the meantime, not only mathematics, but even logic has been formalized. Thus, we may try to leave some tasks to computers: shuffling numbers, finding solutions to differential equations, checking proofs, finding proofs. But whereas checking a supposed proof is relatively easy, finding a proof is rather difficult. Our problem is the following: Do computers prove anything? Or is it just the mathematicians who prove something with the help of the computer?

There are some well-known cases where computers were essential in proving theorems. The four-color theorem is best known. That four colors suffice to color a normal map was conjectured about 1850. In 1976, Appel and Haken completed the proof using 1200 hours computer time in order to check finitely many cases. In the meantime, other mathematicians have found a simpler proof reducing the computer time needed to 12 hours on a workstation. It is thinkable that someone comes

up with an even simpler proof dispensing with the computer altogether. In that case, the computer would have been needed only temporarily. In any case, we would not claim that the *computer* proved the theorem. It was *people* with the help of a very busy tool.

But there are other cases. There are deduction machines (or deduction programs run on computers). They are hoped to find proofs. New proofs. By methods too extensive to be used by man. Until recently, only well-known proofs could be recreated. But now, the *first serious computer proof* has been found. A conjecture from mathematical logic ("every Robbins algebra is Boolean"), about sixty years old, has been proved by a program EQP working with clever trial and error. 'EQP' stands for 'Equational Power', since the program untiringly inserts equations into other equations and then simplifies the ensuing expressions. It had to make about ten thousand insertions, one hundred and five of which built the final proof. This proof could not have been *found* by hand, but it can easily be *checked* by hand! In this case, we may indeed claim that the computer (plus program) found the proof. This does not preclude us from attributing the proof as well to William McCune who devised the program.

People might object. Computers, or programs, or formal systems in general, they might argue, don't *understand* what they are doing. They are just executing formal steps, but they don't know what mathematics is, what logic is about, what truth is, what problems and solutions are, what conjectures, axioms, definitions, theorems are, what truth conservation means, what valid inferences and correct proofs are. It's people, it's programmers, it's men who know and understand all that. Thus, proofs are not computers', but men's.

Who is right? There is no definite answer to this question. It's neither purely empirical nor purely conventional. We have to consider what we *want* it to mean that someone gives or finds a proof. Our (the authors') answer is this: We prefer a terminology in which a computer (plus program) may indeed find a proof. Thus, we claim that computers can prove something and that by now a genuine computer proof has been found.

f) When does a system see something?

As with calculating, we start with a truism: Most men can see. And many other animals can see as well: apes, dogs, mammals and vertebrates in general, but also insects, spiders, octopusses. According to Ernst Mayr, eyes have been invented independently at least 40 times in evolution. They may be very different: lenses, pinholes, pits, stalks, facets, infrared receptors. But they are all called eyes, and all these animals are said to see something.

We could be more demanding. We might ask for special eyes (lenses, e.g.), for special receptors or light-sensitive substances (rhodopsin), for special wavelengths (380 to 760 nanometers, as in man), for special processes (simultaneous processing of stimuli, two-dimensionality), for special achievements (in sensitivity, sharpness, color distinction). But usually, we don't care. We ascribe seeing to all organisms with eyes.

Do artifacts see? Does a camera see something? No. It does not process the information it gets. It just stores it. This applies to photo and film and TV cameras. They don't see. Nor do periscopes, telescopes, microscopes, bioscopes, gyroscopes, nor, for all etymology, any "scopes" whatsoever. It's always humans who *use* such things in order to see. Could robots see? Yes, if they process optical signals adequately. When do they do it adequately? Unfair enough, we leave this question unanswered. Maybe later.

g) What is thinking? Oh pardon: When do we accept a system as thinking? When should we?

Men think. Which other systems do so? Extraterrestrials are supposed to think. (They do it in science fiction.) But even with them, the problem arises: Do they *really* think? Could they not just feign to think, just look like thinking? How would we find out? Do animals think? Of course, there is both evolution and development in thinking. Earthworms don't think. Nor do embryos or newborn children. (Or is even this, after all, not a matter of course?) Finally: Do machines think? If not, could they think, at least in the future, at least in principle? These questions are notoriously difficult, and the difficulties have, as usual, two sources: vagueness of concepts and problems of empirical evidence.

Is performance enough? Alan Turing, in his well-known paper "Computing machinery and intelligence", proposes to replace the question "Can machines think?", which he thinks to be meaningless, by another: If a machine tried to imitate a human being in written communication, could we find out that it is a machine? This has become known as the "Turing test". We may reformulate Turing's proposal as follows: "Could a machine pass the Turing test, at least in the future, at least in principle?"

Turing even made a prediction: "I believe that in about fifty years' time it will be possible to program computers, with a storage capacity of about 10^9, to make them play the imitation game so well that an average interrogator will not have more than 70 per cent chance of making the right identification after five minutes of questioning." According to this prediction, made in 1950, computers should by now pass a Turing test at least for five minutes. It is clear that Turing's prediction has not come true. But suppose it had: Would we be ready to say that computers can and do think? Would we content ourselves with behavior, with performance? If the machine successfully imitates man, would we be ready to attest it thinking?

Our test has a great advantage: We don't need to know what thinking exactly is; we can do without a clear-cut *definition* or explication. All we need and have instead is a nice *criterion.* To be sure, it's rather superficial, behavioristic, black box; but nevertheless, it is *workable.* It is also *plausible* since it ties in with our ascribing mentality to animals. It can even be made quantitative: How many minutes do you need to unmask the machine? But is the Turing test indeed a sufficient criterion for thinking? And even if, is it necessary? Don't we suppose monkeys to think although they haven't got the slightest chance to pass the Turing test?

The answers are not evident. We exhibited examples where *performance* was sufficient: flying, fixing Rubik's cube, seeing. In other cases, we do care for the *mechanism* by which the performance is achieved: calculating, for example. We are generous in some cases and restrictive in others. But why? Nobody knows. Is thinking more like flying or more like calculating? Since there are examples for both, there is no standard way to decide this question, no a priori solution, no induction from a wealth of known cases. Whereas Turing sides with the Turing test, i.e. with performance, Ziff cares for mechanisms. We will have to search deeper, to look for more arguments and for more knowledge.

We take it as uncontroversial that humans think. How do we know? Of course, *I* think. But do *you* think? How do *I* know? Maybe you are an extraterrestrial? Or a robot, cleverly designed? Or both? And vice versa, how could I prove to *you* that *I* am thinking? This situation is symmetric: I have the same problem with you as you have with me. In this case, luckily, we need not restrict our test to performance. We can find out more about us: We are both members of the same species *Homo sapiens*, we have a very long common past, our genes are essentially identical, our brains are similar, we eat the same things, our milieu and education have much in common, we communicate in a common language, our brain processes are essentially isomorphic, etc. All this suggests that we both think. It would be silly to concede thinking to me, but not to you.

But when it comes to other animals, this line of argument becomes less convincing. Once we go down in the phylogenetic tree, the common ground diminishes. And with extraterrestrials or robots it's bound to vanish. Of course, there is always *something* in common: existing in space and time, some cosmic environment, being made of elementary particles, atoms, molecules, following the laws of nature, etc. But what, then, is decisive for thinking, the common traits or the differences? Thus, we have to decide on the criteria by which we attest thinking. If it is not performance alone, is it structure, constituents, mechanisms, algorithms, functions, or what?

h) When is a system conscious?

Maybe later.

B Foundations

B 1 Simulation and realization

May artifacts have mental properties? There are three different answers.

- According to our working assumption (A1c), *every mental property exhibited by a natural system can be exhibited by an artificial system, at least in principle.*

- There is a weaker claim, made by "weak" defenders as well as by "weak" opponents: *Artifacts may have some, but not all mental properties.* There should and could then be mental properties no artifact may ever have.
- According to strong opponents, *no artificial systems will ever have mental properties.*

To be sure, both meaning and truth of these claims depend on the conditions we demand for *having* mental properties. Even so, all these claims are meant to be *factual* or non-analytic: Their correctness should not be derivable from the meaning of the terms. Here, we are not ready to prove or disprove empirical claims concerning brains or artifacts. But we are ready to explicit what it means for a system to *have* a property. First, we have to distinguish between *simulation* and *realization* of a property. And we have to elucidate the hidden assumptions we make when using these terms.

First of all, we will distinguish several *levels.* Usually, our *primary* interest is directed towards the surface properties of a given system. Realization or simulation of the system's surface properties is a simple reproduction of these properties. Let us call this the surface level. But there are more levels.

Take chess as an example. Garri Kasparov is an excellent chess-player, world champion in 1997. Deep Blue is a computer chess program. Run on a fast computer, it beat Kasparov in 1997. Does Deep Blue indeed play chess, or does it just *simulate* playing chess? That the program does not move the pieces, is not essential: We could add a robot and then ask whether Deep Blue plus robot plays chess. Their performance is excellent, even better than the world champion's. But performance is not all there is.

There are deeper levels. On these deeper levels, there *are* differences between Kasparov and Deep Blue, between man and machine. They follow different procedures, different strategies. Kasparov uses some kind of gestalt perception. He recognizes positions and arrangements like a knight fork or a strong center. He considers linear sequences of possible moves, but only a few of them for as many as twenty half-moves or more. Deep Blue, on the other side, uses a *brute force* method: It tries awfully many moves, many more than Kasparov could ever hope to check. It builds huge trees of possible sequences and follows all their branches up to eleven or twelve half-moves and some of them even further. Thus, it checks many possibilities, it evaluates them according to the quality of the position reached, it decides between different moves. Given a special position, both may reach the same decision. This decision, however, is brought about by very different mechanisms.

Regarding such underlying mechanisms or properties, we may repeat our question: Is there a deeper level producing the properties which, in turn, generate the surface behavior? And this question we may iterate again and again. This is shown in our "Zipper"-diagram.

The diagram includes an example: There are several mechanisms to get meals cooked, electric stoves and gas stoves. On the "surface", they do the same thing, namely cooking, and this is what they are made for. But there are deeper levels where they have different properties, different structures, consist of different materials, follow different laws. There are, of course, even more ways of cooking: coal burning, microwaves, volcano rocks. Further examples could be presented: different types of printers (ink, needle, laser), of motors (gasoline, Diesel, hydrogen, electric), of narcotics (chloroform, ether, sledge-hammer).

Given the idea of *levels*, we now try to explicate the terms 'realization' and 'simulation'.

For any given system claimed to realize something or to be a realization of something, we immediately ask: a realization *of what*? Thus 'to realize' is a predicate with two variables, a two-place relation: A system X realizes a property Z.

And a simulation? Again, it is systems which simulate something. But what do they simulate, systems or properties, or both? A system may simulate another one *with respect to a property or several properties.* Hence, 'to simulate' is a three-place predicate: System X simulates system Y with respect to a property Z. But sometimes, or, statistically speaking, even in most cases, we talk elliptically: We let system X simulate system Y without mentioning the relevant property, or we let system X simulate property Z without explicit reference to the system(s) simulated. Deep Blue simulates a chess player (a system) or simulates playing chess (a property).

Obviously, realization is *more* than simulation. The difference is about the same as between origi-

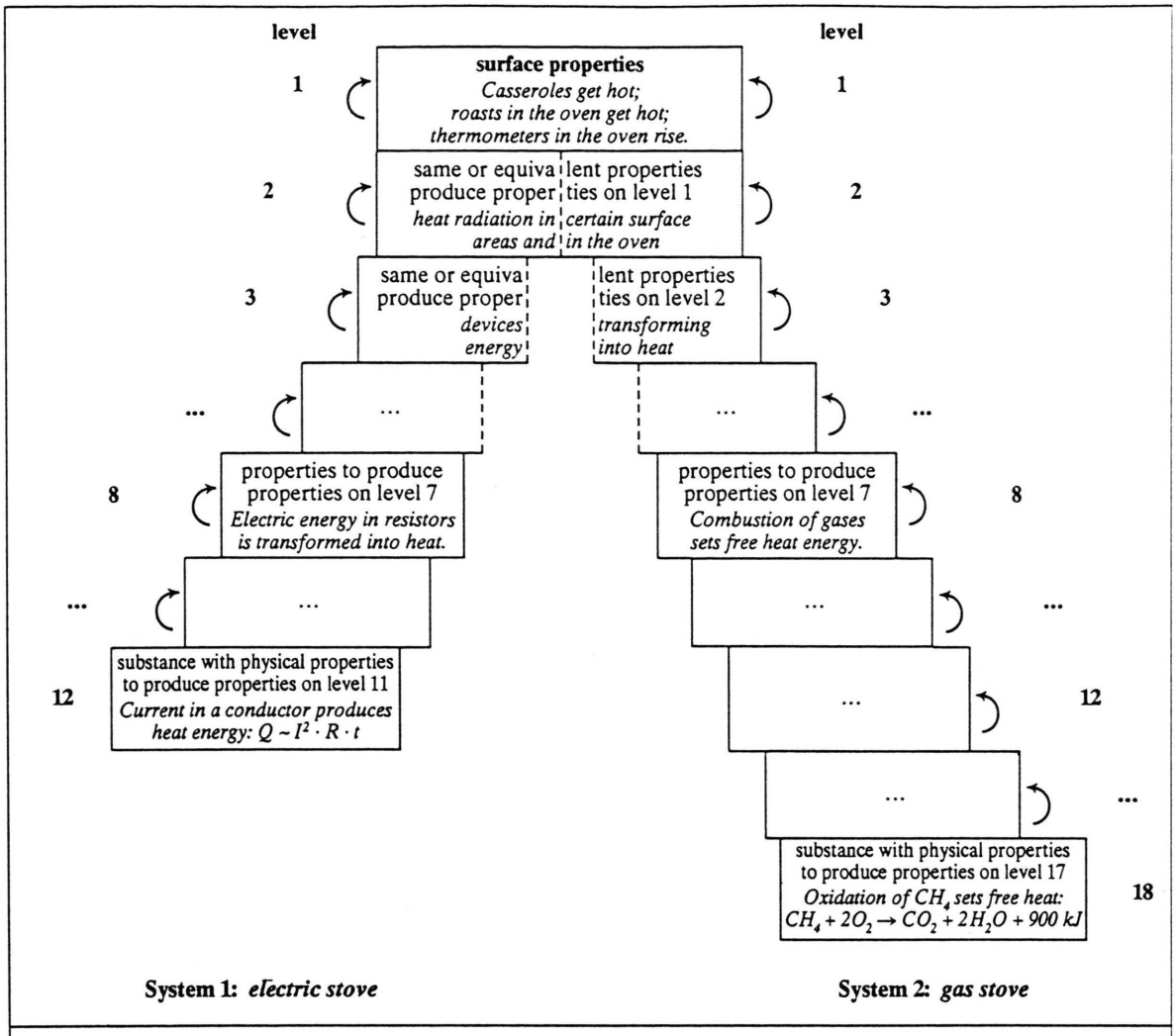

Diagram: **Zipper of realization/simulation.**
Different constituents on lower levels may produce or **realize** equal properties on upper levels.

nal and copy, between authentic and imitated, between genuine and faked. A realization not only *simulates* a property but *has* it. According to Beckermann (1997), a system S *realizes* a property P if and only if S has underlying properties and if, according to the laws of nature, any system having these underlying properties shows the higher property P.

In our stove example, the underlying properties are heat production and heat transport by radiation, conduction or convection. The surface property or behavior is cooking. The higher properties are derivable via laws of nature, first of all via the laws of thermodynamics. Taken this for granted, the gas stove as well as the electric stove realize the *same* surface property on level 1 (or 1 and 2) by *different* processes on deeper levels. As usual, we can take any property on any level as our new property and ask if there are underlying properties realizing it.

But nobody would say that a gas stove simulates an electric stove as a whole or in every respect; it just *realizes* some of its properties such as heating or cooking. And vice versa: The electric stove *real-*

izes some of the gas stove's properties, but not the gas stove as a whole. An electriv stove is not a gas stove.

In the light of our zipper, the difference between realization and simulation is, on any level in consideration, a matter of degree. A simulation realizes *some* but not *all* properties of that level. A *total* simulation (X simulates Y in every respect) is no longer a simulation but a realization. Vice versa, a realization might be seen as an extreme case of simulation. A *partial* realization, then, is not a realization in the strict sense, but a simulation.

Let's consider some further aspects:

- A system that realizes or simulates another one need not necessarily have the same number of functional levels as the system realized or simulated: it may have less or more.
- Given any level, it is often not easy to decide on the next underlying level. Take Gestalt psychology. Why do people tend to complete a line with some hidden part preferably to a straight line? Gestalt psychologists will explain this by appeal to some underlying mechanism, the principle of the good Gestalt. A neuroscientist may claim that this tendency is due to neuronal processes obeying rules of parsimony. In this case, the claim of a Gestalt psychologist is treated as more superficial, more phenomenological, merely descriptive.
- Strictly spoken, no system can simulate *itself* (because it *is* that system). Nor can a system simulate a system of the same *type* all over: A man cannot simulate being a man, and a planet cannot simulate being a planet; but a man can simulate an ape, and a fixed star or a plane may look like or be taken for a planet. But of course, a man can simulate himself in respect to a certain property: You might simulate being drunk or having some kind of disease. People then might call you a pretender or malingerer.
- Normally one system simulates another one, but not vice versa. The simulated system serves as a predecessor, either because it was in existence prior to the simulating system or, more often, because the simulated system is natural and the simulating system is artificial. (Remember the use of the term "natural" for living systems.) Quite often the simulated system is more powerful than the simulating system (although this need not be the case). Combinations of these relations are possible.
- If S_1 simulates S_2 then S_2 must be known: There is no simulation of something which is unknown. In that sense S_2 is *prior* to S_1.
- A simulating system never has all the properties of the simulated system. In using the term 'simulation' you implicitly admit this. In other terms, 'to simulate' means nothing else but 'to imitate, to do as if'. And when using 'as if', you imply that the simulating system equals the simulated system in some, but not in every respect.
- Why did we call the diagram above 'the zipper of realization/simulation'? Because it is necessarily Y-shaped (or Λ-shaped) like a zipper: The two sides of the zipper – its "legs" – converge on the upper levels (upper by drawing, lower by number). And it's equally clear that two different realizations cannot not be identical in every respect; to be different they must differ at least on some level, on a more fundamental one, in their substantial make-up. Therefore, there has to be a transitional stage in which the two systems have equivalent properties.
- Note that our talking of *properties* refers not only to structural properties, but also to physical ones. (See B2.)
- By the way: 'Realization' and 'instantiation' are taken as synonymous. Take red objects: Any red object *has* the property of being red, it *realizes* this property or *instantiates* it.

B 2 Aims of research: explanation

The diagram above and our considerations may be used to distinguish *several types of explanations.* As scientists we are primarily interested in explanations. We dedicate most of our time to find good explanations, and we want to rule out bad or insufficient explanations. There are different types of scientific explanations. Which kinds of explanations are used in the sciences of the mind? What are their qualities? And are there special problems with them?

a) Description and explanation

Given a system we may ask for a *description.* Perhaps we describe the system's properties, its behavior or some of its mechanisms on this level or even the interaction of such mechanisms.

With respect to any description we may further ask: *Why* is that so? By this we ask for an *explanation.* Explanations usually refer to underlying levels. Having an explanation at our disposal we are able to show *how* and *why* the properties previously described obtain.

Description and explanation are *relative* concepts. Once we have achieved an explanation we may take the explanatory components (the *explanans*) as a description by itself and ask for an explanation of the facts described there. Thus we ask for a mechanism on the *next* level. In other words: We want an *explanation* for our previous explanans. Obviously we can repeat this for any explanation given, until we reach the basic concepts and principles of the discipline or even the laws of physics.

b) Explanation of emergent properties

When discussing emergent properties and their explanation it may be useful to know what philosophers mean when they talk about *emergence* (Beckermann *et al.*, 1992, Stephan, 1997). The word was first used by Lewes in 1885, but the idea can be traced back to Mill. According to *Webster's New Encyclopedic Dictionary* 'to emerge' means: 'to come into view, to become known or apparent, to rise from an obscure or inferior condition'. Although this is quite similar to what philosophers mean by emergence, dictionary meanings or etymology do not *determine* the meaning of technical terms. The latter are determined by conventions of a scientific community. So in distinguishing two versions of emergence (weak and strong) we do not state what *nature* is like: It only means that we prepare the tools we want to use in research and discussion.

'Emergence' is a term which can be used for properties or for laws. Usually it is applied to complex systems, systems composed of simple parts and arranged hierarchically. Emergentists are usually monists, holding that there is only one reality consisting of physical elements. This is the position of physical monism. Furthermore they hold that there are systemic properties: properties exhibited by the system, but not by its parts. And they keep to the thesis of mereological supervenience: The properties of the systems are determined by the properties of the elements and their composition. There can be no difference in system properties (on the higher level) without differences on the lower level. And the lower level determines what happens on the higher one. We call this *microdetermination*. Thus, three elements define a *weak version* of emergentism:

(1) physical monism,
(2) existence of systemic properties,
(3) mereological supervenience or microdetermination.

Without doubt such weak emergent properties exist, even in relatively simple physical systems. The fluidity of water or the eigenfrequency of a resonant circuit are weak emergent properties. They can be explained. The explanation of a weak emergent property is a special kind of explanation by the laws of nature. It is more complicated than average because the components do not show the behavior of the entire system.

Strong emergence is weak emergence plus the condition of irreducibility: There could be systemic properties of complex systems in the physical world which are microdetermined by their components and their relations but which are not reducible to the properties and relations of those parts, that is, the high-level properties *cannot be deduced* from the lower level even if we have complete knowledge of the lower level. So we have as a fourth component:

(4) irreducibility.

Are there strong emergent properties in nature? Does the brain exhibit strong emergent properties? What would it mean to reduce those properties to the components and their relations? This depends very much on our concept of reduction. Some philosophers only require that we should be able to explain why the relation of mereological supervenience holds.

A good test to determine if it is possible to explain the relation of supervenience is to look for *macrodetermination*, that is for a causal influence of the macroproperties on microproperties which cannot be explained by the causal interaction of components and their relations. So far, no convincing example of macrocausation has been found.

In the discussion about emergence, other criteria for emergence were proposed, e.g. true novelty and unpredictability. They may be contingent features of strong emergence, but they are not necessary.

c) What are functional explanations?

What do we mean by *function*? Sometimes nothing else is meant but the system's behavior. How does a light switch function? Look, you push here, and the light goes on. You push there, and it goes off. In this case 'function' means 'behavior' or 'input-output-relations'. This case is not problematic; so let us concentrate on a more sophisticated one.

Let us start with some general considerations consulting *Webster's New Encyclopedic Dictionary* again. In mathematics a function is a relationship assigning to some or each element of a given set exactly one element of the same or another set. In everyday life a function may also refer to a professional position or duty *or* to a social ceremony or social gathering. In what follows we ignore those everyday meanings.

If we could rebuild an organism atom by atom (or quark by quark) equivalence is not the problem. But if we construct complex artifacts we must have an idea of what the system is "good for": We must specify its *functions*. We must define the *problems* it is meant to solve. Even here, there are different types of function and, therefore, different types of functional explanations.

The discipline most intensely occupied with functions in the sense of useful performance is evolutionary biology. So we may learn something for artificial systems if we consider the role of functions and functional explanations in biological systems. A function may be

- the action for which a person or thing is specially fitted or used or for which a thing exists,
- one of a group of related actions contributing to a larger action.

Many things have functions in the first sense: Wings are for flying, the heart is for pumping blood, chairs are for sitting, etc. The two paradigm classes of functional devices are biological organs and artifacts. In artifacts the function originates from the intention of the designer. They make appeal to goals and purposes. That is why we call them *teleofunctions*. What about functions of organs? Who has a goal or purpose in their case? That biological organs, structures or behaviors fit the world was for a long time one of the most convincing arguments for the existence of God (God as the great designer) before evolutionary theory. But with the emergence of evolutionary theory there was another explanation at hand: Biological devices fit the world *because* they were selected for their functions. So the existence of systems with teleofunctions is explained by natural selection or other selective processes.

But functions are used by biologists also in our second sense: Studying complex systems with interrelated parts we assign functions to parts or subparts because they do something for the system as a whole. The function of a certain enzyme is its role in the process of oxidative phosphorylation. Those functions are often called Cummins functions after the philosopher who described this kind of functional analysis. We will call them *mereofunction* (from greek *meros*: part).

But there are also functions which have not evolved, were not intended by their designers: Books may function as paper-weights and noses may support glasses. In other words, an object may function *as* something. If a function is defined by what a device can potentially do every device has indefinitely many functions (though not every device can do everything). We call this kind of function *mechanofunction*.

We have now three type of functions (Walter, 1998, 196):

- *Mechanofunction*: Everything a device can do or be used for.
- *Mereofunction*: The function a certain structure fulfills in relation to the system it belongs to.
- *Teleofunction*: The function of a device for which it was selected or designed for.

The study of mechanofunctions and mereofunctions will not tell us much about teleofunctions. Although mereofunctions are probably those functions engineers are mostly occupied with, teleofunctions are the most important to understand natural systems and thus potentially also for building artificial systems equivalent to natural system.

There are at least two reasons: Teleofunctions can only be explained by looking into the history of a system. They explain, at least partly, why certain structures *exist*. Structures may have teleofunctions even if they don't have them as mechanofunctions. A malformed heart has the teleofunction of pumping blood (that is what it is selected for) even though it may not be able to pump blood. We thus may say that an organ is *supposed to* do something because it belongs to a successful

line of descent and is a copy of ancestors which were served by this function in a certain way. The 'supposed to' in this context sounds normative but can be naturalized by referring to a selection history.

Note the analogy to engineering problems. A robot may be *designed* to fulfill a certain function even if it does not succeed. But whereas it is possible to construct failing robots for years, nature is not so generous: If a system is not functioning well it will vanish. This may be due to internal reasons (hearts which don't pump blood will not be reproduced) or because competing systems function a little better and multiply faster. Thus we end up with systems performing certain functions at least to a degree which allows them to survive under competition.

There is another difference between natural and artificial systems. Aircraft engineers may replace steel by titanium almost over night, by clever invention and total reconstruction. By contrast, organisms, depending on viable mutations, may change their biochemical design only if every intermediate step is selectively advantageous. Thus insects cannot replace chitin shield by cartilage, let's say, or by a bone skeleton.

The idea of teleofunctions may give us a heuristic tool to construct artificial systems. Actually this is the case in the field of genetic algorithms: They develop and get selected by performing certain functions for certain systems.

Teleofunctions are important for another reason: They may be at the core of the phenomenon of intentionality. That is, the meaning of a representation may be defined by its teleofunctions. If this is true, meaning cannot be read off the structure of a system but only by looking at its developmental history. Notice that we now speak of development in general and not only of organismic evolution. The reason is simple: If there are selection processes satisfying certain criteria, there may be teleofunctions on shorter time scales than the time scales of evolution. There are, indeed, such selection type theories of brain functioning. So engineers of artificial systems might be well advised to use the concept of teleofunctions not only for discovering features of complex systems or algorithms in order to build them into artificial systems, but rather to build selective mechanisms into the cognitive systems themselves!

If teleofunctions define intentionality in natural systems, this would explain why arguments like Searle's (the intentionality of artificial systems is not *genuine* but only *derived*, see D2c) are true by the very nature of intentionality. Whereas natural systems inherit their functions from predecessors and owe them to the processes of natural selection, artificial systems owe them to the intentions of their engineers. And by definition this cannot be done otherwise!

One last word: In functionalism one is concerned with mental states and with the role other mental states play for them. Thus, a mental state is defined in terms of its relation to other mental states and its effects on behavior. When I think of taking an umbrella with me when leaving my apartment, other mental states are related to this thought: my knowledge about the weather, my preference of being dry to being wet, the effort it will take me not to forget the umbrella etc. Having these mental states as a background my thought of taking an umbrella is the consequence of my previous mental states. Therefore, these previous mental states are the explanation of my actual mental state.

But is this functional explanation in the sense of functionalism a sufficient explanation in our sense? Of course not. Functionalism does not tell the whole story. Clearly, other mental states are in some way responsible for having a special mental state. In our account other mental states are *mereofunctions* for the present mental state. This taken for granted, functionalism is only capable of explaining how different mechanisms (in this case: different mental states) are related to other mental states. This is at best an explanation on level k to describe the mechanisms and consequences on level $k-1$. But level k is certainly not the basic level, and not the substantial one. So functionalist explanations can only serve as an intermediate stage but not as the desired one. Why should we not ask: And what are the underlying mechanisms producing mental states at all, at least in a naturalistic worldview? Before we have not reached the basic laws of physics no explanation is complete and, therefore, the functionalist account is insufficient. In other words, functionalism neglects hardware. It it does that by intention.

d) The structural account or: Is having the same structures enough for having the same physical properties?

Suppose you have two systems made from different substances. Suppose, that from level k up to higher levels (with lower numbers) these systems show the same mechanisms and the same behavior. And suppose that on the last level where the underlying mechanisms are different you are not only able to decribe these different mechanisms but you also find an isomorphism between the elements and between their relations. That means you can structurally map these systems onto each other on this level (or even higher levels). Then, obviously, you are allowed to speak of *structural equivalence.* Sooner or later the thought may arise, that having common structural traits is *sufficient* for having the same mechanism on a certain level and therefore the same properties. Although these mechanisms or properties are admittedly *realized* by distinct substances, and therefore the mechanisms under examination are different, they are identical concerning their structural traits. Once it is clear that a structure alone (on a certain level) is enough to explain the mechanisms on a higher level you are allowed to speak of a *structural explanation* for these mechanisms. But how do you know that it is enough? From a philosophical point of view there are three positions concerning mental properties:

- an extreme bio-chauvinism;
- an extreme structure-chauvinism;
- a mediating position.

This is what the positions state in detail: According to *bio-chauvinism*, you must take into account the substantial realization of a structure. Just as there is only one molecule having a density-anomaly at 4 °C (namely H_20) and just as there is only one element showing hardness 10 (carbon atoms in diamond structure), there is only one physical make-up able to produce consciousness (being made out of "flesh-and-blood").

This is Searle's view. This view is, of course, a possible one, and it should also be possible to find good arguments for it. If you accept it, artificial systems will only be able to *simulate* natural systems because some features of the bio-hardware can never be realized by abiotic devices.

According to *structure-chauvinism*, the physical substances realizing specific structures are of no interest at all. Take the structural equivalence of computers as an example. On which computer you program a certain task, is of no importance; it is the program that matters. More precisely: Computers are all-purpose machines. Assume you have written a program for a given computer to solve a certain task. Now, take a different computer. For this computer you write a second program which makes it work just like the first one. Then you are able to run your first program on this computer.

Obviously, you can take the first program and its structure as a sufficient description of how to solve the given task. Any program having the same structure will fulfill the same task. The way it is bound to the hardware is of no importance. Thus you may say: The structure of the program serves as an *invariant* over all possible realizations in factual computers and therefore it is enough to concentrate on structure alone. Consequently, all programs having the structure of the invariant realize the same structure and therefore the same properties.

This view is very common in approaches to the mind that leave out hardware explicitly. Let us call it the structural or algorithmic approach. This approach to mental properties suffers from an intrinsic weakness: Stating that the brain works algorithmically, at first glance it seems possible in principle to realize mental properties by running algorithms on a different hardware. Especially: If you describe the brain's processing by algorithms, one should assume that having all necessary algorithms at hand, a silicon-based device, able to process all these algorithms, *has* mental properties just as the brain *has* them. The problem: You can describe the algorithmic work of the brain on very different levels. If so, how do you know that you have chosen the right level? Take multiplication as an example: Solving a multiplication task may on a high level be described just as we learn it in school. On a deeper level, an algortihmic description concentrates on the algorithmic processing of certain neuronal modules. On the next deeper level you may have to take single neurons and their biochemical processing into account. On the next deeper level ... – You will find algorithms on all levels. But they do not have to be identical, indeed, they almost never will. Summarized, the

structural approach seems to be possible, but is left with the intrinsic problem of choosing and hitting at the right level of algorithmic description.

Following a *mediating position* we might say that describing and rearranging structures on a medium or algorithmic level alone might not be enough to build artificial systems with the desired mental properties. Searle is right in arguing that the substance of a system *might* be responsible for specific properties not exhibited by simply rearranging certain elements on a higher level of description. But Searle may as well be wrong in assuming that *no* other substance can realize all properties of bio-systems. So, what we have to look for is special elements in special arrangements having the properties of the bio-arrangement. Take as non-biological systems an electric stove and a gas stove: Obviously, no gas stove *is* an electric stove (or has all its properties). And no gas stove *simulates* an electric stove (nor vice versa). But surely, concerning the essential properties of stoves, a gas stove *realizes* a stove.

C The Case for Artificial Mentality

Our working assumption reads: Every mental property exhibited by a natural system can be realized by an artificial system, at least in principle.

Pondering over arguments supporting this hypothesis, we should distinguish truth and benefit. So, why could one think the assumption to be true? And why would it be useful to think so?

a) Analogy from supposed brain mechanisms

Mental properties (faculties, states, processes) are properties of special organic systems: brains, central nervous systems, organisms. To simplify matters, let's take brains. It could be that brains work algorithmically, at least on some level. And it's just on algorithms that some artifacts, namely computers, are so terribly good. Being essentially universal Turing machines, they may perform or instantiate *any* algorithm. True, there are space and time restrictions. But these count as well for (or against) brains. Thus, it is perfectly thinkable that computers may equal brains in their achievements. Try hard, and you should succeed.

This argument may be countered in two ways. You either deny that mental properties are properties of brains. (You might claim that they rather come with persons, or minds, or subjects.) Or you deny that brains work algorithmically. (To be sure, you can deny both.) For these arguments see below.

b) Successes so far

In fact, artifacts realize many surprising properties. They achieve things nobody would have expected. Which reads: We did it, so why should we not do the rest? What is more, many predictions on the *un*abilities of machines have indeed been falsified. Conclusion: You skeptics were wrong on so many things, so you must be wrong on all things.

How good is this argument? While the proponent points to a wealth of successes, the opponent will hint at so many failures. Thus, the situation might seem symmetric. But is it? How are we to divide the burden of proof? Who has to come up with better evidences or more arguments? Logically speaking, what makes the trouble is the mixed universal-existential claims. Ambitious adherents of artificial mentality would claim (see our working assumption): For *every* mental trait, *there is* (thinkable) an artifact realizing this very trait. This claim can neither be proved nor be refuted. As a universal statement it is unprovable: Even if artifacts show *some* mental traits, we might come up with another trait that is still elusive. And as an existential statement it is, at the same time, irrefutable: Even if *so far* no artifact shows mental traits, we might still claim that *future* devices will. Thus it is not testable at all.

But our opponent is not better off. In contrast to our proponent he will claim: No, *there are* mental traits which *all* artifacts (will or must) lack. And confronted with artifacts showing mental traits, he can always challenge the proponent by naming traits unrealized so far, whereupon he will be countered by more promises for the future. Thus, as logic might have taught us, our opponent's claim is not testable either.

There is a way out. We could be less ambitious. We should not talk about *all* mental properties, but only claim that *some* of them are realizable by artifacts. The weaker version of our working assumption then reads: *Some* mental traits can be realized by *some* artifacts. Obviously this is a purely existential claim. Although it cannot be refuted it may be proved. It is, however, up to the proponents to prove it.

It is here that we might encounter another serious trouble. Whenever in the eyes of the proponent an artifact has a mental property, his critic might easily retort: He either denies the mentality of the said property ("but calculating is not a genuine mental activity"), or he questions the artifact's *really* having this mental property ("but this is not calculating in the *true* sense"). This evasive strategy can be iterated indefinitely. The only way out seems to be that both partners agree *beforehand* what is to count as a real mental property and when they are ready to ascribe it to an artifact. But even then, someone might change his mind on what is mental. And who can afford to deny learning to his opponent?

c) What exactly do you mean?

Another provocative argument reads: Tell me *exactly* what computers can't do, and I'll write you a program doing it. Thus, John von Neumann says: "Anything that can be completely and unambiguously put into words is ipso facto realizable by a suitable finite neural network." (He thought however that *complexity* would forever forbid such a complete description, hence prevent engineers from building an artificial brain.) (McCorduck, 1979, 65) This claim is tricky since it shifts the burden of the argument to the critic. The idea behind is that specifying in detail a natural (or organismic, or human, or mental) property already amounts to giving a pertinent algorithm which would realize that property. But this is not true. We may list all chess rules and specify precisely what it means to win, and still do not know how to win. Up to now, nobody has a winning strategy for chess. A detailed description of the behavior wanted does not always amount to an algorithm producing this behavior.

Most of these arguments are not really pro; they just stress that there is no convincing contra.

d) Heuristic value: Could it be useful to suppose that artifacts may have mental states?

If we think computers, or engineering, or science in general, to be ridiculous, there is no point in finding mental states in artifacts or in supplying them with mental properties. There is, then, no point in research at all nor in scientific heuristics. If, however, we are eager to search into the inner workings of nature or to find out about the prospects of artifacts, for instance computers, how will we proceed? How much does our approach depend on our convictions? Is it important whether we believe in artificial mentality?

Of course, there will be *some* effect. It is the difference between optimism and pessimism. If you hope to find something, you will look more carefully, improve methods, convince others, try harder. If you have no hope you won't try. But if you don't try you won't find anything. Hope makes a difference.

If we believe that artifacts can have mental capacities, we will try to build them. Does that mean that a disbeliever will turn to poetry? Not necessarily. While being convinced that artifacts will lack mentality forever, *he might still try* in order to find out how far he can go, where the obstacles lie and in what fundamental ways artifacts are different from natural systems, from humans, or organisms, or brains. For curious men *both results are interesting*. There is, then, as far as research strategies are concerned, no sharp dividing line between believers and disbelievers in artificial mentality.

Besides, most scientists are skeptics. They know they are fallible. They don't consider but one possible outcome. They think about the odds that they are wrong. Knowing that they might be wrong they will try to find out if in fact they are wrong in order to detect and to correct their mistakes. At least they *should* do all that. Thus, the difference between the two parties is not as big as might have seemed. Believe in artificial mentality or not, you will try both ways.

D The Case Against: Supposed Objections

D 1 A preliminary classification

Is thinking really like flying? Is it only a matter of convention whether we attribute thinking to a system? There seem to exist some serious obstacles to do so. After all, thinking is not the only problem. What about seeing? Seeing something red? Having pain? To be in love? It seems that humans have properties that artificial systems will never have. Philosophers' favourites are consciousness and language. But what about animals? Aren't they conscious too? Maybe they are. But how could we ever come to know? Don't we only

observe their *behavior*? To attribute consciousness to them because they behave *as if* they were conscious seems precarious. A simple vehicle may be programmed to follow lights and to avoid walls. It behaves *as if* it were intentionally searching light, *as if* it didn't want to bump into the wall. Obviously, our use of language is largely metaphorical. Here at least we are not ready to attribute mental states, much less consciousness. We claim to know better. But how do we identify other humans as being conscious? This is the "other-minds problem". Their behavior alone cannot be sufficient to justify our judgement since we do not accept this criterion for machines either. So why do we think humans to be conscious? Because we know ourselves to be so by introspection, because we register their verbal reports about their own mental life and because we know that they are about constructed as we are. Still, we can never *prove* them to be conscious or to have mental states at all.

Instead of dwelling on the honourable, but very clumsy problems of consciousness and introspection, we will try to pin down the core problems of why artificial systems are said not to have any mental life at all. For the moment we are not interested in high-end questions like: Can machines be self-conscious or have free will? (Free will is, after all, at issue even for *humans*.) Rather, we are interested in *arguments*. *Why* should artifacts not be able to exhibit mentality at all. Turning down high-order properties, people even doubt that artifacts are able to experience, to believe, to know, to think, or to understand anything! Let's discuss some of their arguments.

For a start, let's sort the arguments against artificial mentality into some categories (see also McCorduck, 1989, 127).

- *Just-no* claims: Mentality is typical of and exclusive for higher organisms, possibly for humans. Artifacts just cannot have mental properties. Period. (This is not an argument, but a statement.)
- It would be *desastrous* to learn that we are not the only systems with mentality. (To be read: I don't like the idea; therefore it must be wrong. – This is no argument either, although it may be quite convincing.)
- *Examples are missing:* So far, no artifact shows mentality. Whether it is possible at all remains to be seen. (Read: The question cannot be decided yet. We don't even know whether, when and how it can be decided, much less do we know the result. Let's go to work and see what else we can find out. – This is a pragmatic argument and honorable for that. Not all people can always discuss on what is possible. But it can be a part-time job or a part-people job. As a theoretical argument, however, it is weak: We cannot conclude from what there is to what there is possible.)
- *Ethical considerations:* Given that artificial mentality is impossible, it would be a waste of time and money to build it. Given that it is possible and in our reach, should we really try? First, we don't foresee the outcome, and the outcome might be bad. Are there not enough examples of detrimental consequences? Should we not postpone research, until we have a better judgement? (This is an interesting problem, but we won't dwell on it. We shall discuss whether artificial mentality is possible and not whether it is desirable.)

If something is impossible, there must be insuperable difficulties. We now list such *supposed* difficulties.

- Insuperable *differences in physical components (for computers: in hardware):* One or several of the following traits are decisive, essential, necessary, constitutive for mentality, but cannot be owned by artifacts:
  - Organisms are made of biomolecules, cells, neurons, tissues, organs. No artifact is. Were we to build artifacts from such components, they might indeed show mental properties. (Searle)
  - The working of brains relies on *quantum effects*. As long as we cannot use these in our artifacts, they won't have mentality. (Penrose)
  - Mentality (consciousness, intelligence, emotions) needs, besides everything else, *high complexity*. Brains are just *too* complex to be imitated by artifacts.
- Insuperable *differences in origin and interaction*:
  - Human beings are outcomes of biological *evolution* under strong selection pressures. We would have to repeat evolution in order to build artifacts with mentality. This takes millions, if not billions of years.
  - Human beings are outcomes of lifelong *development and learning*.

- Human beings do have contact with the *real world*. This contact is decisive for learning, intelligence, emotions, mentality. Especially:
- Human beings are *social* beings. They live with and learn from each other.

• Insuperable *differences not attributable to physical components, origin or interaction:*
- Human beings understand *meanings*. Artifacts may equal or even outdo humans on abstract, algorithmic, computational, logical levels, but they cannot *understand* anything. To them, there is only syntax, no semantics, no reference to something real or outside, no intentionality.
- In problem solving, humans use *intuition*. Intuition is, nearly by definition, not definable in algorithmic terms and not realizable by artifacts. (Dreyfus)

D 2 The main arguments

a) The argument from emotions (folk psychology)

Lay people do not hesitate to attribute some kind of *intelligence* to artificial devices. They readily accept computers or robots outdoing humans in cognitive domains. But is seems obvious to them that artificial systems – at least if they are not made from flesh and blood – do not have *emotions*. Correspondingly, people with "cold" rationality showing no emotional sensitivity are called *inhuman*. Emotions seem to be an indispensable constituent of human mentality, but not amenable to artifacts.

Answer: How seriously should we take this popular argument? Quite. Emotions *are* a problem for the sciences of the artificial. Until very recently they were neglected although there are good reasons not to do so. Evolutionary theory has shown that emotions have important functions for organisms (Nesse, 1989): They are automatic routines for the fast appraisal of situations; they play a central role in guiding organisms; they motivate (Dörner and Hille, 1995); they solve the commitment problem (Frank, 1988); they are involved in decision making (Damasio, 1994); and they have cognitive functions (Power and Dalgleish, 1996).

Why, then, have emotions been neglected after all? Because they are very difficult to explore. In psychology there are even problems to *define* emotions properly. Although the neurosciences have gained a lot of detailed knowledge about the cerebral structures and transmitters underlying emotions there is as yet no good explanation of the bipolar nature of emotions, much less a unified theory.

These problems with emotions in psychology and neuroscience do perhaps explain why modelling of emotions in artificial intelligence is the exception, not the rule. But emotion and cognition are linked. As long as artificial systems have no motives, no curiosity, no fear, no affiliation, no hunger and no thirst, they will not become really intelligent. Why should they? In fact, why should a system be intelligent? Intelligence is needed when a system, be it human, animal or machine, tries to reach a goal with a minimum of energy (energy is scanty), to go straightforward (detours require time and energy), or when it tries to avoid conflicts between motives or with fellows (conflicts are costly).

The modelling of emotions is indispensable if we want to build artificial organisms which have goals and plans, are able to stick to them if appropriate, keep their priorities in order.

Many people would have it that it is even impossible to *simulate* emotions. Recent experiments falsify this (Hille, 1997). In robotics the importance of the representation of goals and motives is now widely acknowledged.

b) Qualia argument (Nagel, Jackson, Bieri, Chalmers)

Whatever emotions might be, there certainly is some *feeling* with them. To feel something is a specific *subjective* experience. Although some feelings are quite diffuse and introspectively impenetrable, others have distinct characteristics. They have a specific quality, a *quale* (plural: *qualia*): A sharp pain is a pain, and nobody will tell me that I am in error about my pain. Pain is something between emotions and perception. But also non-emotional perceptions do have a qualitative character: *It is something like* to see red. And more complex mental states do have a subjective quality, too. Whereas we know what it is like to be in pain, to see red, or to be a father, it is very difficult, if not impossible, to know what it is like to be a bat. What is more, even if we knew all about the anatomy, physiology, biochemistry and computational

architecture of bats we would still have no idea what it is like to be a bat.

Likewise, even if we knew everything about color perception, the physical (or physiological) story would not tell us whether a human being's subjective experience of red is like mine. Maybe he or she experiences some kind of blue, but consistently calls it red? I can *imagine* a person who has the same cerebral color systems as I have, calls everything red that I call red, and everything blue what I call blue, but who experiences those colors just interchanged (inverted qualia). How could I ever get to know that? Well, how do I know that (s)he experiences anything at all (absent qualia)? Couldn't (s)he be a zombie?

This argument is not restricted to higher-order mental states. The explanation of feeling pain or of seeing red is a problem already "down" in animal physiology (the difference being that we *know* that we are in pain or see red whereas for animals we have to infer it.)

The qualia argument has been considered as unsurmountable for a physical theory of mental states (Nagel, 1974; Jackson, 1986; Bieri, 1995; Chalmers, 1996). Its general structure is: It is logically or conceptually *possible* (*imaginable*) that systems possess *all* structural (physical) and functional properties of conscious systems without being conscious, that is, without having the phenomenal experience of, say, seeing red. Therefore, structural and functional properties do not *necessitate* conscious experience. In other words: Given all natural laws, physical structures and functions of a system, phenomenal conscious experience does not follow. Thus, it must be thought of as something additional, intrinsic, fundamental, and irreducible. Qualia and phenomenal conscious experience are candidates for what we have called (in) strong emergent properties: they are not only systemic, but irreducible properties.

For an *answer*, see the end of the next argument.

c) Intentionality restricted to organismic systems (Searle)

Besides the explanation and, possibly, the artificial making of qualitative experience (call it raw consciousness), there is another general feature of human mental life which seems to resist understanding and which is denied to artifacts: intentionality. Intentionality refers to the fact that human mental states have content, that they point to an object, that they *mean* something. Again, intentionality is not thought of as a necessarily complex, advanced or high-order mental state, but as a *basic* feature of human mental state. There is a well-known objection which argues that intentionality or meaning is still a mystery: the Chinese-room argument (Searle, 1980).

Suppose you are enclosed in a room with all kinds of books or files. Through a slot you get written questions in Chinese. You don't speak or read or understand any Chinese. But in your files you find advices how to proceed with all kinds and combinations of Chinese signs. You follow some kind of algorithm. At the end of this procedure you are able to produce, through the slot, a written answer in Chinese signs. You still don't understand Chinese, not even the answer you produce. Thus, Searle claims, understanding hovers over and above all algorithmic procedures. Since computers work by algorithms, they cannot *understand* anything, they cannot have intentionality.

Answer: Although this argument is closely related to the argument from non-algorithmicity (e) we will discuss it here. First of all, a general consideration has to be taken into account: Obviously, both the qualia argument (b) and the present argument from intentionality are based on *thought experiments.* Kathleen Wilkes (1993) criticizes the use of such thought experiments in the philosophy of mind as pure armchair-philosophy. For scientific purposes, she prefers *science fact* to *science fiction.* Dennett uses the term "intuition pump" in order to disqualify the role of thought experiments in debates about mental events. Why? Here, thought experiments are not used to *illustrate* certain points, but as *arguments.* But are they *good* arguments? – Thought experiments are experiments with boundary conditions which are not and, usually, cannot be realized. In order to get results you introduce conditions, devices or creatures that do not exist in real life, demons for example. Or you use your intuition to judge the result. Case studies (Buschlinger, 1993) reveal that in both cases you may be misled. Even in physics this danger is common as shown by Maxwell's famous demon: Introducing the demon implies hidden assumptions that destroy the argumentative power of the thought experiment. Because in almost every case such

hidden assumptions go unnoticed, it is difficult to decide on the quality of the argument. Hence, we have to be careful in taking a thought experiment as an argument unless we are sure that we have taken all necessary (but sometimes hidden) assumptions into account.

Second, and as a direct objection to the Chinese-room thought experiment taken as an argument: Of course, the single person inside the Chinese room does not understand anything. She is – just like the CPU of a computer – only one part of the comprehensive system. The slot belongs to the system, too, as well as the input-ouput devices and the files and advices how to process the signals. Regarding this, the system *as a whole* might understand Chinese, whereas none of its parts does. At least, Searle's claim concerning the system is not as convincing as it is for the person inside.

d) Time as essential in natural systems
(van Gelder, Port)

Artificial organisms are thought of as instantiations of Turing machines. For Turing machines, time is irrelevant. Not so for natural systems: If they are not fast enough they fall down or get eaten. Thus, the computational demand for natural brains is not just to solve problems by any number of steps, but to solve them *in real time.* It should not come as a surprise if this constraint had favored a certain type of cognition in evolution. But "What could cognition be if not computation?" (van Gelder, 1995). This question may sound strange to people working in artificial intelligence. Can there be any non-computational cognition? Yes, there can.

The two contenders so far as theories of cognition are classical *computationalism* (GOFAI: good old fashioned artificial intelligence) and *connectionism.* Both postulate the existence of representations (symbolic or distributed) being worked on by specific algorithms. "Non-computational cognition", then, must postulate cognitive processes which do not depend on such representations processed by algorithms on the usual level.

Dynamical systems theory, or dynamizism, claims to be the third contender. Dynamical systems are state-determined systems whose behavior is governed by differential equations. Their variables are evolving continuously and simultaneously and at any time mutually determining each other's evolution. A dynamical system has a number of component behaviors which must be time-dependent, deterministic, generally complex, low-dimensional, and intimately linked. Such a system is directly coupled to its environment and influenced by constraints effective in real time.

For dynamizism, cognition is the evolving behavior of dynamical systems directly coupled to their environment. Cognition is not time-independent computation but state space evolution of a complex system in real time. The apt metaphor is not the Turing machine, but the Watt governor (Fliehkraftregler) in a steam engine. Although a Watt governor can be simulated by a Turing machine, it is, on the working level, not a Turing machine. A Watt governor consists of real parts being embedded in an environment and having their behavior determined by natural laws *without representing* these properties, *without computing* anything on representations of anything. Likewise, a cognitive system is not a computer. It is not a brain either, isolated and encapsulated. It is rather an integrated system embracing nervous system, body, sense organs and environment. A cognitive system is not a discrete sequential manipulator of static representational structures, but rather a structure of changing and mutually interacting parts. It does not interact with other parts of the world by exchanging information, it *coevolves* with them.

To the ear of an engineer this may sound all very metaphorical and a little conjuring. But there is a growing bulk of work done by using dynamical systems theory. In Port and van Gelder (1995), many concrete examples are given, ranging from locomotion via development, visual perception, audition phonology, and syntax, to language. Most advanced seems to be developmental psychology, especially the development of movement (Thelen and Smith, 1994). Instead of going into details let's look how the dynamical systems approach could be used as an argument against the possibility of artificial mentality.

Answer: Two comments might be appropriate. First: Port and van Gelder do not think their approach to be incompatible with connectionism. Connectionist nets may be understood as a very special kind of dynamical systems. Thus, the dy-

namizist approach is the more general framework, encompassing other models. Second: Dynamizists do not claim that artificial systems could never be dynamical systems. They rather stress that natural systems are essentially dynamical systems and that we should adjust our modelling to that fact if our goal is to build equivalent systems. In a nutshell, their argument is: "It is unlikely that it will be possible to reproduce the kind of intelligent capacities that are exhibited by natural cognitive systems without also reproducing their basic noncomputational architecture." (Port and van Gelder, 1995, 4). We conclude that their argument is not an argument against the possibility of artifacts with mental life, but rather a methodological challenge to reframe our kind of modelling.

e) Non-algorithmicity (Searle, Lucas, Penrose)

This is the claim: Computers work algorithmically, brains don't. Therefore computers can not be equivalent to brains.

But how on earth do we *know* that brains rely on more than algorithms? There are several supposed arguments. Let's study three paradigm cases.

- The first is Searle's Chinese room argument presented and answered in c.
- The second argument relies on Gödel's incompleteness theorem and was brought forward by Lucas (1961). According to Gödel, for every formal system (consistent and rich enough to contain number theory), there are statements which can be formulated in S, yet cannot be proved (nor refuted) in S, but nevertheless are true and can be *proved* to be true by richer means. Now Lucas claims that every computer (or any artifact working algorithmically) can be looked at as a formal system (consistent and containing arithmetics). Following Gödel, then, for *every* computer C there will be propositions which are true and can be seen to be true by men, but cannot be proved by this individual computer C. Thus, men and computers must be *different*. Since men can do something computers can't, men are even *superior* to computers in that respect.
  Answer: This argument is defective. We don't doubt that computers comply with the presuppositions of this theorem. Nor do we dispute Gödel's proof. Thus, we will concede that Gödel's conclusion applies to computers: Every computer misses some true propositions men can prove. But is this any different with man, i.e. vice versa? Brains either work by algorithms, or they don't. If they do, then Gödel's theorem applies to them as well. Then every human being or brain will miss some true propositions some computers might be able to prove. If not, then Gödel's theorem does not apply and nothing is left for an argument. Since Lucas wanted to *show* that brains do not work algorithmically, it's of no use to take this claim as a premise.
  In addition, it's by no means clear whether we always find the proof asked for. It might be too complex for a finite mind (as it is for a finite computer). Thus, the situation is completely symmetric. No difference is found, and no superiority either.
- The third argument is brought forward by Penrose (1989, 1994). Being a mathematician and a theoretical physicist, he looks at the problems and the problem solving methods of his disciplines. And he gets the *impression* that scientists don't work algorithmically. He just cannot *imagine* that what he and his fellow scientists do can be caught in an algorithm.
  Answer: This is not an argument, it's more or less a confession.

f) Externalism (Putnam, Burge)

Traditional brain sciences suppose that both natural and artificial organisms can be explained by looking at their *internal* structure. Identitist philosophers as well as most neuroscientists claim that mental states are identical with brain states: "The astonishing hypothesis is that 'You', your joys and your sorrows, your memories and your ambitions, your sense of personal identity and free will, are in fact no more than the behavior of a vast assembly of nerve cells and their associated molecules. As Lewis Carroll's Alice might have phrased it: 'You're nothing but a pack of neurons.'" (Crick, 1994, 3).

There is a family of arguments against this view based on the theory of *externalism*. Externalism claims that my propositional attitudes cannot be characterized without reference to objects and properties in the world – to my environment:

"Thoughts just ain't in the head." (Putnam, 1975, 223; Burge, 1979) Again, most of these arguments are based on thought experiments (cf. D2c). Imagine an exact replica of John: Zohn. Zohn lives on a twin earth, a planet exactly like our earth with one single exception: There, the substance we call water is not composed of H_2O but of XYZ. Thus, whenever the thoughts of Fred on earth are about H_2O, the thoughts of Zohn on twin earth are about XYZ. So the content of a thought does not always depend on internal structure alone but also on external factors.

To put this argument in a more general frame: Representations are essentially *relational* properties; they cannot be individuated without reference to external elements. If we want to understand why a structure may represent something at all, syntactical structure is not sufficient. Hence it is not sufficient to rebuild isomorphisms between two kinds of organisms or their brains but the isomorphisms must hold as well for their relations to external factors.

Answer: Is this an argument against artificial minds? We do not think so. Rather, it is an argument which shows that *content* is not an intrinsic property of the brain, be it natural or artificial. Content rather lives on the interaction between a cognitive system and its environment. The idea of embedded cognition has recently gained attention in considerations on artificial life (Steels and Brooks, 1995) as well as in philosophy (Bermúdez *et al.*, 1995). It is more and more acknowledged that the *embedding* of the brain is a constitutive feature of cognitive systems. This is not only true for detached environments, but also for its immediate surrounding, its body.

g) History and development (Millikan, Dretske)

Natural organisms have a natural history, artificial organisms don't. Natural organisms develop: They get born, grow, are educated and change, sometimes quite drastically. Artificial organisms normally are ready-made. How far does mentality depend on history or development? And how are these observations turned into arguments against the mentality of artifacts?

In B2c we discussed teleofunctions. Teleofunctions (also called *proper* functions) are intrinsically historical. If they are essential for mentality, artificial mentality becomes questionable. Indeed, there are claims that teleofunctions are at the core of intentionality. The theory of teleosemantics (Millikan, 1984; 1993; Papineau, 1987; 1993; Godfrey-Smith, 1994) being rather complicated, let's put it in a nutshell: Why do certain states in the visual system of the frog represent insects and not fast moving dots or shadows? Because this is the interpretation they were selected for! Now, just as teleofunctions of evolved *organs* depend on their evolutionary history, the intentionality of *thoughts* depends on and is constituted by the history and evolution of the brain processes realizing mental states.

At first sight this might look as a striking argument against the mentality of artificial systems. But taking a closer look we see that this is not the case. According to teleosemantics, every system developing under selection pressures will exhibit teleofunctions. Hence other devices, e.g. language devices and artifacts, may have teleofunctions, too. The reason why bottle-openers exist is not exhausted by the fact that they have the mechanofunction of opening bottles, but by their being *designed to* open bottles. Their ability to open bottles plays an indispensable role in the explanation of why bottle-openers do exist. Whereas natural systems do have a natural design (explained by the theory of natural selection), artifacts have an intended design (explained by the intentions of their designers).

If teleofunctions are at the core of intentionality we come to a surprising conclusion. Searle's claim that the intentionality of artificial systems is not *genuine* but only *derived* is true by the very nature of intentionality. The claim that only biological systems have true intentionality must be read as: Biological systems do have intentionality because they have teleofunctions explained by the theory of natural selection. And artificial systems have intentionality explained by the intentional design of their engineers.

But does this theory not strengthen Searle's argument? How could artifacts ever come to have the kind of intentionality humans have? Is it not obvious that humans are able to think in ways very different from the ways evolution has shaped them to think?

Answer: This objection can be countered. Teleosemantics isn't restricted to evolutionary functions.

It can be used to develop a "neurosemantics" (Kurthen, 1992). According to Millikan (1984), for a teleofunction to emerge it is necessary that (i) there is a family of devices, (ii) these devices get copied, (iii) the performance of the function by some members of the family *explains* the existence of actual devices, at least in part. Thus teleofunctions can be established in other processes than evolution. Dretske (1988) for example has developed a teleosemantic theory depending on development, e.g. the learning history of individuals.

The extension to shorter time scales need not stop here. If the criteria just mentioned apply to brain processes there may be selection processes in the range of minutes down to seconds. There are even theories of brain functioning (Edelman, 1987; Changeux and Dehaene, 1989; 1995; Sporns and Tonini, 1994; Calvin, 1996) meeting those criteria; they are characterized as selection type theories (Darden and Cain, 1989). They use principles of natural selection to explain brain functioning. Apparently there are selection processes in the brain, e.g. when categorizing stimuli or when selecting from alternatives for action. This may explain how the brain generates new meanings (Walter, 1998).

To summarize: Intentionality depends on the history of the devices in different time ranges. This is relevant in three ways. First, it explains why artifacts have *derived* intentionality. Second, it gives us a conceptual tool for the study of natural systems; understanding their design would enable us to build analogous artifacts. Third, it gives us a clue how to build artifacts generating new meanings: They should have intrinsic selective mechanisms which allow them, by interaction with the environment, to acquire new teleofunctions and thus to become more autonomous and independent of the intentions of their human designers.

D 3 Kinds of Objections

It might be helpful to classify the preceding arguments according to the kind of pertinent questions:

- Empirical: Do the objections point to gaps in our empirical knowledge which might be overcome by the advancement of science?
- Methodological: Do the objections criticize a certain methodology and propose some new methods or heuristics?
- Epistemological: Do the objections point to problems caused by our limited epistemic access to certain phenomena?
- Conceptual: Do the objections give a fundamental argument against the equivalence between natural and artificial systems regardless of empirical, methodological or epistemological progress?

The objection of emotion is dependent on empirical constraints: We might discover much more about emotions than we dare to dream of. The objections of time, of externalism, and of history, have strong methodological as well as epistemological aspects. They point to some hitherto neglected aspects and provide new models or even new mathematical tools incorporating those aspects into the building of artificial minds. The objections concerning qualia, intentionality and non-algorithmicity are mainly conceptual. That's why they are at the core of philosophical debates.

These objections are not neatly separated: The emotion and qualia problems share the feeling problem; the intentionality argument refers to the non-algorithmicity of semantics; the objections of externalism, history and time are related by the fact that natural systems are embodied and coupled to their environment.

Why is it useful to distinguish between these kinds of objections? First of all it gives us an idea of what the objection is about. In addition, it may give us a clue how those objections can be countered.

Those which are mainly *empirical* or *methodological* can be answered by the 'just-a-problem' strategy. They give no reasons against the possibility of artificial minds but rather point to serious difficulties in their design. To designate those objections as 'just-a-problem' is not to say that they are unimportant or easy to answer. It means that they are rather constructive than destructive. They point to blind spots or omissions. They are tools for error elimination and insofar standard elements of research.

Epistemological aspects point to limits of our cognitive, especially our self-reflective, capacities. If history is essential for understanding intentionality we may never uncover the true meaning of something. The best thing to do is to invent plausible stories matching with the possible course of events. Presumably this is what we do when we

engage in folk psychological discourse (Dennett, 1987). In the case of qualia, epistemological considerations also play an important part. In order to solve this problem it is necessary to find out how a system represents itself and how far it is cognitively penetrable.

For the *conceptual* aspects it is important to have clear ideas or, even better, theories on the problems under discussion: What is an algorithm? What do we mean by saying that something (thinking, computing) is algorithmic or not? When is a system experiencing a quale? Are there unanalyzable mental predicates, are there respectively irreducible mental properties? What does it mean to mean something? What kind of complexity do we require of a system in order to attribute a property to it? Are there basic forms of mental activities? We must explicate our terms, sort our arguments, and be clear on the level of our discussion. In this paper, we have tried to do something about it.

Acknowledgements

We thank our colleagues Hermann Wagner and Achim Stephan for their helpful suggestions.

Anderson A. R. (1964) (ed.), Minds and Machines. Prentice-Hall, Englewood Cliffs.

Beckermann A. (1997), Property Physicalism, Reduction and Realization. In: Mindscapes: Philosophy, Science, and the Mind (Carrier M. and Machamer P. K., eds.). University Press, Pittsburg, 303–321.

Beckermann A., Flohr H. and Kim J. (1992) (eds.), Emergence or Reduction. Springer, Berlin, New York.

Bermúdez J. L., Marcel. A. and N. Eilan (1995) (eds.), The Body and the Self. MIT Press, Cambridge.

Bieri P. (1995), Why Is Consciousness Puzzling? In: Metzinger.

Braitenberg V. (1984), Vehicles. MIT Press, Cambridge. (= Künstliche Wesen. Verhalten kybernetischer Vehikel. Vieweg, Braunschweig (1986). = Vehikel. Experimente mit kybernetischen Wesen. Rowohlt, Reinbek (1993).)

Burge T. (1979), Individualism and the mental. Midwest Studies in Philosophy **4**, 73–121.

Buschlinger W. (1993), Denk-Kapriolen? Königshausen & Neumann, Würzburg.

Calvin W. H. (1996), The Cerebral Code. MIT Press, Cambridge.

Chalmers D. (1996), The Conscious Mind. Oxford University Press, Oxford.

Changeux J.-P. and Dehaene S. (1989), Neuronal models of cognitive functions. Cognition **33**, 63–109.

Changeux J.-P. and Dehaene S. (1995), Neuronal models of cognitive functions associated with the prefrontal cortex. In: Neurobiology of Decision-Making (Damasio A. R. *et al.*, eds.). Springer, Heidelberg, New York, 125–144.

Crick F. H. C. (1994), The Astonishing Hypothesis. Scribner's & Sons, New York.

Damasio A. R. (1994), Descartes' Error. Putnam's Son, New York.

Darden L. and Cain J. A. (1989), Selection type theories. Philos. Sci. **56**, 106–129.

Dennett D. C. (1987), The Intentional Stance. MIT Press, Cambridge.

Dörner D. and Hille K. (1995), Artificial souls: motivated emotional robots. Proceedings of the International Conference on Systems, Man and Cybernetics. Vancouver, Vol. **4**, 3828–3832.

Dretske F. (1988), Explaining Behavior. MIT Press, Cambridge.

Edelman G. (1987), Neural Darwinism. The Theory of Neural Group Selection. Basic Books, New York.

Frank R. (1988), Passions within Reason. Norton, New York, London.

Godfrey-Smith P. (1994), A Modern History Theory of Functions. Nous **28**, 344–362.

Hille K. (1997), Die künstliche Seele. Deutscher Universitäts Verlag, Wiesbaden.

Jackson F. (1986), What Mary didn't know. Journal of Philosophy **83**, 291–5.

Kurthen M. (1992), Neurosemantik. Enke, Stuttgart.

Lucas J. R. (1961), Minds, machines and Gödel. Philosophy **36**, 112–127. Reprinted in Anderson (1964).

McCorduck P. (1979), Machines who Think. Freeman, San Francisco.

Metzinger T. (1995) (ed.), Conscious Experience. Imprint Academic, Thorverton (dt: Bewußtsein. Paderborn, Schöningh).

Millikan R. G. (1984), Language, Thought and Other Biological Categories. MIT Press, Cambridge.

Millikan R. G. (1993), White Queen Psychology. MIT Press, Cambridge.

Nagel T. (1974), What Is It Like to Be a Bat. Philos. Rev. **83**, 435–50.

Nesse R. M. (1989), Evolutionary explanations of emotions. Hum. Nat. **1**, 261–289.

Papineau D. (1987), Reality and Representation. Blackwell, Oxford.

Papineau D. (1993), Philosophical Naturalism. Blackwell, Oxford.

Penrose R. (1989), The Emperor's New Mind. Oxford University Press.
Penrose R. (1994), Shadows of the Mind. Oxford University Press.
Port R. and van Gelder T. (1995) (eds.), Mind as Motion. MIT Press, Cambridge.
Power M. and Dalgleish T. (1996), Cognition and Emotion. Psychology Press, Hove, East Sussex.
Putnam H. (1975), The meaning of meaning. In: Mind, Language and Reality (Putnam H.). Cambridge University Press, 215–271 (deutsch: Die Bedeutung von Bedeutung. Klostermann, Frankfurt (1979)).
Putnam H. (1979), Robots: machines or artificially created life? In: Mind, Language and Reality (Putnam H.). Philosophical Papers, Vol. **2**. Cambridge, University Press, 386–407.
Searle J. (1980), Minds, brains and programs. The Behav. Brain Sci. **3**, 417–424.
Searle J. (1983), Intentionality.
Sporns O., Tonini G. (1994), Selectionism and the Brain. Academic Press.
Steels L. and Brooks R. (1995) (eds.), The Artificial Life Root to Artificial Intelligence: Building Embodied, Situated Agents. Lawrence Erlbaum, Hillsdale.
Stephan A. (1997), Emergenz. Habilitationschrift. Universität Friedericana zu Karlsruhe.
Stieve H. (1995), Limits of natural science: brain research and computers. Z Naturforsch. **50c**, 317–336.
Thelen L. B. and Smith L. B. (1994), A dynamical system approach to the development of cognition and action. MIT Press, Cambridge.
van Gelder T. (1995), What might cognition be if not computation? Journal of Philosophy **91**, 345–381.
Walter H. (1998), Neurophilosophie der Willensfreiheit. Schöningh, Paderborn.
Wilkes K. V. (1993), Real People. Oxford, University Press.