Wolfgang Lenzen

Intrinsic Intentionality

1 Introduction

Intentionality is a key-notion in Searle's philosophical thinking. It plays a dominant role not only in many of his articles on the philosophy of language and the philosophy of mind, but it also figures as the title of one of his books. For reasons of space I cannot discuss all the issues which are dealt with there [Searle 1983]. These issues cover such diverse (though interrelated) topics as the Intentionality of mental states; the Intentionality of perception and action; Intentional causation; Intentionality of language and communication in general plus Intentionality of indexical expressions, of natural kind terms, and of proper names in particular. Thus I will here not touch upon the derived Intentionality of language, i.e. on the systematic meaning or function which the intentions of a person have for the purposes of linguistic and nonlinguistic communication, because these issues rather belong to another section of this symposium on "Speech acts and Intentionality". Instead I want to focus on the basic distinction between genuine or "intrinsic" Intentionality on the one hand and derived or "metaphorical" Intentionality on the other hand and try to shed some light on the question:

> (***II***) *What are the necessary and sufficient conditions for intrinsic Intentionality?*

This question lies at the very heart of contemporary research in the philosophy of mind. Of course nobody should expect me to offer a definite answer here and now. Actually no philosopher alone might ever answer it in a complete and satisfactory way. Question (***II***) only outlines the rough direction of an interdisciplinary research program which will have to be carried out primarily by neuroscientists, biologists, and psychologists, but to which philosophers can also contribute in a not altogether unimportant way.

One preparatory task which modern philosophy should solve within this larger enterprise of cognitive science is to remove the stumbling blocks that other philosophers have laid in the way toward a scientific understanding of how the brain "produces" the mind. Searle himself has taken some important steps in this direction, e.g., by arguing that the traditional "mind-body problem" is only a metaphysical pseudo-problem which arises from an uncritical use of "a seventeenth century conception of [the body and the mind]" [Searle 1984, 6]. At the very beginning of *The Rediscovery of the Mind*, Searle presents his philosophical position of a so-called "biological naturalism" as follows:

> The famous mind-body problem, the source of so much controversy over the past two millenia, has a simple solution. [...] Mental phenomena are caused by neurophysiological processes in the brain and are themselves features of the brain. [Searle 1992, 1]

This view provides a *partial* answer to question (***II***) by maintaining that one decisive condition for intrinsic Intentionality consists in being a living, biological organism of a certain highly-developed kind. Already in a 1984 paper on "Intentionality and Its Place in Nature", Searle maintained that:

> Intentional mental phenomena are part of our natural biological life history. Feeling thirsty, having visual experiences, having desires, fears and

> expectations, are all as much a part of a person's biological life history as breathing, digesting, and sleeping. Intentional phenomena, like other biological phenomena, are real intrinsic features of certain biological organisms in the same way that mytosis, meiosis, and the secretion of bile are real intrinsic features of certain biological organisms. [Searle 1984, 5]

However, what biology (in conjunction with neurology) teaches us in the first instance is only that being a living organism of a certain complex, highly-developed kind just is *sufficient* for displaying certain Intentional phenomena. It does not, however, strictly prove that being a living organism of a certain complex, highly-developed kind would also be *necessary*. In order to avoid any possible misunderstandings, let me add that I personally do believe that - as far as *biological organisms* are concerned - having a structurally complex central nervous system and brain *is* necessary for displaying Intentionality. In particular I am convinced that bacteria, fungi, flowers and trees have no Intentionality at all; and the same will probably also hold of lower animals which possess only a very primitive brain. Accordingly I was very surprised to see that one of the most-debated contemporary authors in the philosophy of mind, David Chalmers [1996, ch. 8] seriously entertains sort of a panpsychistic view according to which not only flowers and trees but also inanimate things such as rocks and electrons are supposed to have some kind of conscious experience (and hence Intentionality). Anyway, Searle's biological naturalism, i.e. the assertion

> (***BN***) Only a living being of a certain complex form will ever have Intentionality,

is certainly not *a priori* true and thus it remains in need of further argumentation. From the fact that so far in nature we have observed Intentionality only in connection with living organisms of a certain kind, it does not logically follow that it is impossible that elsewhere in nature Intentionality might be realized in altogether different systems.

A second philosophical contribution to the interdisciplinary task of answering question (***II***) consists in a systematic analysis of the large - and rather inhomogeneous - field of Intentional states. This might lead to a subdivision of all mental states into the most important subclasses of, say, perceptions, feelings, beliefs, desires, etc. Accordingly one can focus on the somewhat easier task of finding answers to the more specific questions:

> (***IP***) What are the necessary and sufficient conditions for intrinsic *perceptions*?
>
> (***IF***) What are the necessary and sufficient conditions for intrinsic *feelings*?
>
> (***IB***) What are the necessary and sufficient conditions for intrinsic *beliefs*?
>
> (***ID***) What are the necessary and sufficient conditions for intrinsic *desires*?

etc. An important further question might be formulated as follows:

> (***IUL***) What are the necessary and sufficient conditions for an intrinsic *understanding of language*?

Searle has dealt with this issue especially in the context of the controversy if, and under which circumstances, *computers* might be supposed to understand a natural language.

2 The "Chinese Room"

In a 1980 paper on "Minds, Brains, and Programs", Searle put forward the famous thought-experiment of the "Chinese room" to show that a computer(program), even it answers Chinese questions in an apparently intelligent way, does not really *understand* the language. The starting point of his investigation is the "script" program developed by Schank and Abelson. This program knows the typical elements of certain events like celebrating a birthday party or having dinner at a restaurant in so far as it is able to answer questions concerning details that have not been mentioned explicitly in the respective "story". This is quite noticeable since the facts or data given in the answer do not deductively follow from the information provided by the story; rather the program compares the latter data with the corresponding "script" and infers the answers by way of so-called default conclusions. Searle's "Gedankenexperiment" now runs as follows:

> Suppose that I'm locked in a room and suppose that I'm given a large batch of Chinese writing. Suppose furthermore, as is indeed the case, that I know no Chinese either written or spoken, and that I'm not even confident that I could recognize Chinese writing as Chinese writing distinct from, say, Japanese writing or meaningless squiggles. Now suppose further that after this first batch of Chinese writing, I am given a second batch of Chinese script together with a set of rules for correlating the second batch with the first batch. The rules are in English [...] . They enable me to correlate one set of formal symbols with another set of formal symbols, and all that 'formal' means here is that I can identify the symbols entirely by their shapes. Now suppose also that I am given a third batch of Chinese symbols together with some instructions, again in English, that enable me to correlate elements of this third batch with the first two batches, and these rules instruct me how I am to give back certain Chinese symbols with certain sorts of shapes in response to certain sorts of shapes given me in the third batch. Unknown to me, the people who are giving me all these symbols call the first batch 'a script', they call the second batch a 'story', and they call the third batch 'questions'. Furthermore they call the symbols I give them back [...] 'answers to the questions', and the set of rules in English that they gave me they call 'the program'. [...] Suppose that after a while I get so good at following the instructions for manipulating the Chinese symbols [...] that from the external point of view - that is, from the point of view of somebody outside the room in which I am locked - my answers to the questions are indistinguishable from those of native Chinese speakers. Nobody looking at my answers can tell that I don't speak a word of Chinese [although ...] I produce the answers by manipulating uninterpreted formal symbols. As far as the Chinese is concerned, I simply behave like a computer; I perform computational operations on formally specified elements. [...] I have inputs and outputs that are indistinguishable from those of the native Chinese speaker, and I can have any formal program you like, but I still understand nothing. Schank's computer for the same reasons understands nothing of any stories whether in Chinese, English, or whatever. [Searle 1980: 284-285]

I take it for granted - and it is meanwhile granted also by prominent advocates of "Artificial Intelligence" such as Roger Schank himself - that a "normal" computer(program) will at best help *humans* to understand a foreign language such as Chinese, while the computer *itself* does not understand the language, nor does it have any other mental state. Furthermore, I largely subscribe to Searle's analysis according to which the computer's failure to understand (the

semantics of) a natural language is basically due to the fact that the program can do nothing else but manipulating formal symbols in a purely *syntactic* way. However, some of Searle's more general conclusions drawn from the above thought experiment seem to be somewhat questionable. For a critical re-examination let us distinguish between two kinds of computers. Computers in the *narrow* sense (such as an ordinary desktop PC) receive all their inputs from keyboards or similar devices operated by human beings. In contrast, computers in the *wide* sense, e.g. *robots*, are equipped with some sort of "interface" which enables them to get certain information "by themselves". Given this distinction, it will be argued below that:

- No computer in the narrow sense will ever be able to pass the Turing test. As applied to the capacity of understanding Chinese, this test consists in a dialogue by means of which a test person must find out whether she is communicating with a computer or with a native Chinese speaker. According to the original proposal in [Turing 1950, 42] the questions and answers may deal with "almost any one of the fields of human endeavor". However, even if the computer is an "expert" in any field you like, its "knowledge" is bound to remain of mere "encyclopedic" character and it will never embrace any of the innumerable "*hic-et-nunc*" facts which can only be known by *perception*.
- Searle's attempt to explain why humans - unlike computers - have Intentional states, seems to be mistaken in at least one respect. Intentionality is not (or at least not entirely) a matter of *biochemistry*. Searle [1980, 299] believes that only organic brains can be "causally capable of producing perception, action, understanding, learning and other mental phenomena", while the "artificial brain" of a robot fails to have this "causal power". However, understanding natural language is not necessarily dependent on any particular stuff. Just like an ordinary computer, so also a brain *alone* has no mental states and no real understanding either. Whether consisting of organic or made of inorganic material, an *isolated* "information-processing system" is bound to remain a merely syntactical "information-processing system" which has no access to the semantics of natural language, and which therefore lacks any kind of intrinsic Intentionality. Intentionality can be ascribed only to an *entire organism*, e.g. a human being or some other higher developed animal, and perhaps also to the "artificial organism" of a robot.
- The so-called "Robot Reply" maintains that if future technique should succeed in constructing a robot which is able to perceive (and to interact with) its environment, and if this machine would be run by a computer with a powerful program to process natural language in such a way that it passes the Turing test, say, for Chinese, then this robot really understands Chinese. Searle's attempt to refute the Robot Reply has failed, but this failure does not by itself support Dennett's contrary claim that a robot with consciousness (or Intentionality) really is possible.

3 The Turing Test

In an oft-quoted paper of [1950], Alan M. Turing suggested to let the question whether a "computing machinery" is able to think be answered by a behaviorist "imitation test". This test consists in a series of questions and answers (e.g., via keyboard and monitor) by which one must find out whether one is communicating "only" with a machine or with a real person. Searle's Chinese room argument was meant to refute the adequacy of this Turing test. However, a *refutation* of the Turing test could only be spoken of if it might plausibly be assumed that a purely formal computer program such as Schank's "script" really is able to pass this test. However, this assumption turns out to be unfounded. Unfortunately, Turing himself did not explain in detail which questions from which fields of knowledge should be

admitted as elements of the imitation test. He mentioned only a few examples from mathematics, chess, and poetry, but he seems to have had the idea of admitting almost every realm of science, art and human cognition, including, e.g., the writing of sonnets. However, one need not resort to such esoteric fields as poetry in order to make clear that a purely formal computer program will never be able to pass the Turing test.

During the past decades many so-called "expert-systems" have been developed for various fields of knowledge ranging from computer programs for railroad or airport schedules to medical diagnosis programs (such as "Mycin") which helps the doctor to find the germs of infectious diseases and to recommend an appropriate therapy. In principle computers may become "experts" for any field of science in the sense that they can get every available information or data from arbitrary domains as, say, music, geography, soccer, traffic-accidents and what not. Now, as has already been pointed out by other philosophers, the requirements of a *general* Turing test are much more demanding than giving "expert" answers to questions from one specific field only. Furthermore, discussions of the so-called "frame problem" have shown that computer programs normally have much too little "world knowledge" which is required for giving appropriate answers (or even intelligent solutions) to simple questions and problems of everyday life. Let me here point to another weakness of computer programs which appears not to have been sufficiently recognized in the literature, namely, each "knowledge" of a computer (in the narrow sense) is bound to remain restricted to *encyclopedic knowledge*. Let me illustrate this by means of an example.

A comprehensive encyclopedia (or a textbook of botany) will contain many facts about roses. For instance, the encyclopedia informs us that the name 'rose' derives from the Latin word 'rosa' which itself leans on the Greek 'rhodon'; the encyclopedia classifies roses into the species of tea-rose (*rosa odorata*) and Chinese or Bengal rose (*rosa chinensis*); the encyclopedia tells us that roses ask for a deep soil with sufficient magnesium and potassium; and so on. These are the kinds of facts which a computer evidently may "know". Moreover, if the computer also has a program for botanic questions formulated in English, it may be able to answer the question 'What are the colors of roses?' by an output like this: 'Typically roses are red; but there are also yellow, white, pink and black roses'. However, no computer might ever be able to answer the much more elementary question 'What's the color of *this* rose?' The reason being that computers (in the narrow sense) *cannot see*!

This trivial example illustrates a fundamental limitation of computers in the narrow sense. They have no access to the large field of non-encyclopedic knowledge, no access to the many *hic-et-nunc* facts such as the fact that there is now a yellow rose standing in front of me. Of course, a computer might *afterwards* be fed with the information that at a particular date *d* a yellow rose was standing in front of W.L.. Every *hic-et-nunc* fact can in principle be *afterwards* transformed into an item of encyclopedic knowledge and then find entrance into the computer's data-base. However, the computer will never know this fact *hic-et-nunc*: here and now. This very limitation, i.e. the inability to answer here and now questions about the infinitely many *hic-et-nunc*-facts of everyday life, seems to be a sufficient reason to predict that no computer in the narrow sense will ever be able to pass the Turing test.

Searle's thought experiment wants to make us believe that a person sitting in the Chinese room might give correct and satisfying answers to arbitrary questions by simply following certain instructions how to handle formal, meaningless symbols. However, in this way not even most elementary questions such as, e.g., 'How is the weather outside?' or 'Where and when were you born?' might ever be answered. Even if the program has stored a complete collection of meteorological data for all places in the world, and a complete collection of the

birthdays of the entire population of the world, it might never infer the correct answers to the above questions because they contain (explicit or implicit) *indexical expressions* referring to a particular time, place, and person. Although the program might give correct answers to corresponding "eternal" questions like 'How is the weather in place *p* at time *t*?', or 'Where and when was *N.N.* born?', in virtue of its property of being a *purely formal program* it could never know the relevant *hic-et-nunc*-facts that the first question was asked at time *t* in place *p* or that the second question refers to the particular person *N.N.* who is currently doing the job in the Chinese room.

4 Perception and Understanding

As was rightly stressed by Searle [1980: 303], all the "information" processed by a computer (in the narrow sense) is only a sequence of formal symbols without any interpretation and without any meaning: "The computer, to repeat, has a syntax but no semantics. [...] All the computer has is more symbols". Because the encyclopedic "knowledge" of a computer has no perceptional basis, it is bound to remain purely *formal* and thus becomes comparable to "information" of the following kind:

> There are three kossels each of which mike a horky mopp. The whumsy kossel and the drumsy kossel mike the kelly mopp while the drilly kossel is miking a mopp that is not miked by the drumsy kossel.

The task of this logic-exercise is to infer by way of formal deduction certain conclusions of the type 'The so-and-so kossel is miking the so-and-so mopp'. But such purely formal deductions lead only to purely formal conclusions which do not tell us, e.g., *what a kossel is* or *how a mopp looks like* or what the relation of miking consists in. The apparently descriptive names and predicates quoted in the logic-exercise are in fact meaningless and might as well be replaced by arbitrary other terms.

Now the descriptive terms contained in the "knowledge"-base of a computer program are in exactly the same sense purely formal, without genuine sense or meaning, and without perceptional relation to the real world. Even if the computer's data bank contains the information that roses are red, the computer does not know in a practically relevant sense how roses look like or what color the color red is. The computer would never be able to say that *this* is a rose or that *that* is red. More generally, computers cannot really be said to *think* or to *understand* anything in the fields of knowledge which they are assumed to be experts for.

An advocate of Artificial Intelligence (AI) will presumably reply that in order for a system (whether a living being or a machine) to be able to *think*, it is not strictly necessary that the system is able to *see*. I agree! Similarly, in order for a system to be able to think, it is not strictly necessary that the system is able to hear; and so on for the other types of perception. Yet, if the system can *neither* see *nor* hear *nor* feel *nor* touch *nor* smell *nor* taste; in other words, if the system has no perceptional access at all to the real world, then the system does not really understand ordinary language. However large its encyclopedic "knowledge" may be, it is bound to remain purely formal, purely syntactic and without reference to the real world; hence without meaning or semantics.

Let me dwell on this point a little further. To understand an ordinary, descriptive language *L* does not require that one be able to give the correct answer to every question formulated in *L*. Understanding does not presuppose omniscience. To understand a language *L* also does not require that one knows every word or every concept of *L*. Understanding does not presuppose

linguistic perfection. However, to understand a language *L* requires not only a basic knowledge of *L*'s syntax and grammar but also a basic knowledge of its semantics, i.e. in particular the knowledge of the reference of at least *some* names and the knowledge of the reference of at least *some* concepts. Here it will not suffice that the system provides for an immanent conceptual network containing definitions for certain expressions in terms of others. Thus, e.g., it will not suffice to know that roses are flowers and that red is a color and that there are other flowers besides roses and other colors besides red. It must furthermore be warranted that the conceptual network is somewhere *tied up with reality*. The system has to know the extension of some singular terms not only in a linguistic, but also in an *ostensive* or *deictic* way. It must be able to *identify* the referents of some singular terms and it must be able to determine the extension of some predicates in an ostensive way, saying, e.g., upon appropriate occasions 'This is a rose' or 'That is red'. Computers in the narrow sense evidently lack such reference to the real world. Therefore a program's "knowledge" of the fact that Chimborazo is the highest peak in Ecuador "means" to the computer just as much or as little as the fact that the drumsy kossel is best in miking horky mopps. Ignoring the subtleties of modal language which sometimes refers to non-real, possible worlds, one may say: The real world is the semantics of natural language. If a system has no perceptional access to the real world, then it remains ignorant of this semantics and therefore it has no real understanding. One might even be tempted to generalize this idea into the neo-empiricist slogan: *No thinking without perception*.

5 Is Intrinsic Intentionality a matter of Biochemistry?

Among several other authors, D. Dennett [1981: 54] has argued that upon neurophysiological analysis of the human brain, our own way of "information processing" might also be characterized as a purely formal manipulation of "symbols". Therefore the verdict of the Chinese room applies to the human brain, too, which hence also is nothing but a syntactic machine with no access to the semantics of a natural language. In a 1990 contribution to the magazine *Scientific American* Searle tried to refute this objection by defining the crucial difference between a computer and a (biological) brain as follows:

> The most obvious difference is that the processes that define something as a computer - computational processes - are completely independent of any reference to a specific type of hardware implementation. One could in principle make a computer out of old beer-cans strung together with wires and powered by windmills. But when it comes to brains, although science is largely ignorant of how brains function to produce mental states, one is struck by the extreme specificity of the anatomy and the physiology. Where some understanding exists of how brain processes produce mental phenomena - for example, pain, thirst, vision, smell - it is clear that specific neurobiological processes are involved. [...] Brains are specific biological organs, and their specific biochemical properties enable them to cause consciousness and other sorts of mental phenomena. [...] All mental phenomena, then, are caused by neurophysiological processes in the brain. Hence,
>
> Axiom 4: *Brains cause mind*." [Searle 1990: 22/3]

The latter quotation suggests the idea that a brain "causally produces" the mind in much the same way as, e.g., the liver secretes bile, the pancreas secretes insulin, and the hypophysis secretes hormones. From this conception it would easily follow that mind can neither be produced by biological organs other than the brain nor by the inorganic "electronic brain" of a

computer. However, as Searle himself came to admit elsewhere, such a conception of the brain having the causal power to "secrete the mind" would be "really too crude". In particular it should be observed that *brains alone* cannot properly be said to "cause" consciousness or other mental states. It is not the brain but rather the entire organism to which mental states such as pain and thirst may be attributed; and, again, it is not the brain but rather the entire living being which can properly be said to have mental capacities such as seeing, hearing, feeling or smelling. As regards the presumable "inner life" of an isolated human brain, it is interesting to note how medical experts describe the state of a so-called "complete locked-in syndrome" which sometimes happens after a hemorrhage in the brain stem:

> In such a case the connecting systems and the centers of the cerebrum and the higher brain stem are not injured. But they cannot get any information from outside. [...] The situation basically resembles [...] a "brain in the vat". [...] The head [...] is only metaphorically the place of consciousness; if the physical world outside is cut off, the brain does not become a pure inner world but it experiences the loss of all relations [...] The isolated brain is no longer a bearer of personality but rather a bearer of madness.

A "locked-in" brain may perhaps still have some rudimentary mental capacities; in particular one might perhaps attribute to it some kind of thinking similar to what we experience in dreams. However, if a human brain had been cut off from the outer world from the very beginning of its existence, then it surely never had developed any kind of understanding or any other mental capacity. It ought to be evident that a permanently isolated human brain cannot think any more than a computer. Therefore the main reason why humans can think while computers cannot is not primarily a matter of biochemistry but rather a matter of having access to the outer world. Perception is the first and foremost prerequisite for thinking and understanding. Without perceptional access to the real world, each symbol-manipulating device - whether biological or not - is bound to remain a purely syntactical machine without genuine understanding.

6 The "Robot Reply"

The preceding explanations should have made clear that perception is a *necessary* condition for any system in order to understand natural language. From this it does not follow, however, that perception would also be *sufficient* for this particular form of Intentional states. Actually, in his critique of the so-called Robot Reply, Searle [1980: 293-4] tried to show by means of a variant of the "Chinese room" argument that perceptual and causal access to the real world would not help any robot to gain Intentionality. However, his argument contains some serious flaws. Searle starts by summarizing the main idea of the Robot Reply as follows:

> Suppose we put a computer inside a robot, and this computer would not just take in formal symbols as input and give out formal symbols as output, but rather it would actually operate the robot in such a way that the robot does something very much like perceiving, walking, moving about, hammering nails, eating, drinking - anything you like. The robot would, for example, have a television camera attached to it that enabled it to see, it would have arms and legs that enabled it to act, and all of this would be controlled by its computer brain. Such a robot would, unlike Schank's computer, have genuine understanding and mental states.

After these explanations Searle praised the proponents of the Robot Reply for having recognized that "cognition is not solely a matter of formal symbol manipulation" but presupposes "a set of causal relations with the outside world". Nevertheless he went on to maintain that "the addition of such »perceptual« and »motor« capacities adds nothing by way of understanding [...] to Schank's original program". Perceptive (and causal) connectedness to the real work does not suffice to endow a robot with Intentionality:

> Suppose that instead of the computer inside the robot, you put me inside the room and you give me again, as in the original Chinese case, more Chinese symbols with more instructions in English for matching Chinese symbols to Chinese symbols and feeding back Chinese symbols to the outside. Suppose unknown to me, some of the Chinese symbols that come to me come from a television camera attached to the robot, and other Chinese symbols that I am giving out serve to make the motors inside the robot move the robot's legs or arms. It is important to emphasize that all I am doing is manipulating formal symbols: I know none of these other facts. I am receiving "information" from the robot's "perceptual" apparatus, and I am giving out "instructions" to its motor apparatus without knowing either of these facts. I am the robot's homunculus, but unlike the traditional homunculus, I don't know what's going on. I don't understand anything except the rules for symbol manipulation. Now in this case I want to say that the robot has no intentional states at all; it is simply moving about as a result of its electrical wiring and its program.

For a closer evaluation of this argument it should first be noted that in the earlier quotation Searle had apparently taken the central premise of the Robot Reply for granted. This premise assumes that the robot is really able to *perceive* his environment - where the word 'perceive' is written without any quotation marks and thus may be assumed to have its ordinary meaning. Later on, however, Searle tacitly begins to weaken this assumption by reformulating it in such a way that the computer is only assumed to "operate the robot in such a way that the robot does something *very much like* perceiving". Still a little bit later Searle grants to the robot only the capacity of some sort of *pseudo*-perceiving as expressed by the word "perceptual" in quotation marks. And at the end of the argument the robot's ability of "perceiving" has entirely disappeared. Its movements are no longer taken to be causally produced by preceding perceptions but, as Searle describes the situation, the robot "is simply moving about as a result of its electrical wiring and its program". From a (methodo-)logical point of view, however, this argumentation must be rejected as inconclusive. The core idea of the Robot Reply consists in the following *hypothetical* or *conditional* assertion:

> (***RR***) *If* one would succeed in constructing a robot which is able to perceive and to interact with its environment, and if this machine would be run by a computer with a powerful program which processes natural language in such a way that it passes the Turing test, say, for Chinese, *then* the robot would really understand Chinese.

In his attempt to refute this If-then-thesis, Searle first starts by taking the assumption of the if-clause for granted. But in the course of his argument he tacitly gives up this premise by categorically claiming that a robot will never be able to perceive (and to interact with) its environment. Even if this claim could be supported by further evidence, the truth of the *conditional* thesis (***RR***) would remain untouched. After all, Searle himself had admitted elsewhere that perception itself represents some kind of mental activity. Hence any system which is able to perceive thereby already establishes some kind of Intentionality.

As regards the truth or falsity of the crucial antecedent proposition itself, Searle appears to argue that a robot will never satisfy the necessary conditions for (intrinsic) perception because all its activities are bound to consist in nothing else but "manipulating uninterpreted formal symbols". However, unlike in the case of computers in the narrow sense, this seems no longer true for a computer in the larger sense, i.e. a robot. For if such a machine would really be able to describe the results of its observation in well-formed sentences of a natural language, then these linguistic expressions could no longer be disqualified as only formal symbols devoid of any meaning. Although the *program* running the machine can be characterized as purely formal, a linguistic output such as 'There is a yellow rose standing in front of me' now evidently refers to a specific object in the real world, hence it has a linguistic *reference* and thus it is no longer purely syntactical.

Furthermore, the above-quoted picture of what appears to go on in a robot's "mind" might equally well be taken as a description of our own brain's activities, i.e., also a human brain might be said to do nothing else but "manipulating uninterpreted formal symbols". Searle does affirm, though, that - in contrast to the robot's homunculus - a human homunculus is *conscious* of what he does and "knows what is going on". But this is only a metaphorical claim and no substantial argument. As a matter of fact, neither in the case of robots nor in the case of human beings anything like a homunculus really exists. The manipulation of symbols in a computer is not achieved by a homunculus but by the computer(program) itself. Similarly, the "manipulation of symbols" in a human brain is not achieved by a homunculus but by the brain itself. However, one cannot say that the human brain - in contrast to the robot's "brain" - is aware of these manipulations and "knows what is going on". As was already stressed above, consciousness may never be attributed to the brain alone but only to the entire (e.g., human) organism. If my brain is "receiving information from my perceptual apparatus, and is giving out instructions to my motor apparatus", it is only *me* but not my *brain* who "knows what is going on". Thus Searle's attempted refutation of the Robot Reply reduces to the claim that a robot's purported "perceptions" are not genuine perceptions because they are not accompanied by *consciousness*. This claim, however, stands in need of further argumentation.

6 Silicon Brains

In his 1993 book *The Rediscovery of the Mind*, Searle described another thought-experiment to show that, even if the robot *behaves as if* it perceives its environment and as if it understands ordinary language, it will never really have consciousness or intrinsic Intentionality:

> Imagine that your brain starts to deteriorate in such a way that you are slowly going blind. Imagine that the desperate doctors, anxious to alleviate your condition, try any method to restore your vision. As a last resort, they try plugging silicon chips into your visual cortex. Imagine that to your amazement and theirs, it turns out that the silicon chips restore your vision to its normal state. Now imagine further that your brain, depressingly, continues to deteriorate and the doctors continue to implant more silicon chips [...]: in the end, we imagine that your brain is entirely replaced by silicon chips [...]. In such a situation there would be various possibilities. One logical possibility, not to be excluded on any a priori grounds alone, is surely this: you continue to have all the sorts of thoughts, experiences, memories, etc., that you had previously; the sequence of your mental life remains unaffected. [...]

> I hasten to add that I don't for a moment think that such a thing is even remotely empirically possible [...]. But that is an empirical claim on my part. It is not something that we could establish a priori. So the thought experiment remains valid as a statement of logical or conceptual possibility.
>
> [...] A second possibility, also not to be excluded on any a priori grounds, is this: as the silicon is progressively implanted into your dwindling brain, you find that the area of your conscious experience is shrinking, but that this shows no effect on your external behavior. You find, to your total amazement, that you are indeed losing control of your external behavior. You find, for example, that when the doctors test your vision, you hear them say, 'We are holding up a red object in front of you; please tell us what you see.' You want to cry out, 'I can't see anything. I'm going totally blind.' But you hear your voice saying in a way that is completely out of your control, 'I see a red object in front of me.' If we carry out this thought experiment to the limit, we [...] imagine that your conscious experience slowly shrinks to nothing, while your externally observable behavior remains the same. [...]
>
> Now consider a third variation. In this case, we imagine that the progressive implantation of the silicon chips produces no change in your mental life, but you are progressively more and more unable to put your thoughts, feelings, and intentions into action. In this case, we imagine that your thoughts, feelings, experiences, memories, etc., remain intact, but your observable behavior slowly reduces to total paralysis. Eventually you suffer from total paralysis, even though your mental life is unchanged. So in this case, you might hear the doctors saying, 'The silicon chips are able to maintain heartbeat, respiration, and other vital processes, but the patient is obviously brain dead. [...]' Now in this case, you would know that they are totally mistaken. That is, you would want to shout out, 'No, I'm still conscious! I perceive everything going on around me. It's just that I can't make any physical movement. I've become totally paralyzed.'
>
> The point of these three variations on the thought experiment is to illustrate the *causal* relationships between brain processes, mental processes, and externally observable behavior. [...] It seems to me there is a number of lessons to be learned. The most important is that they illustrate something about the relationship between mind and behavior. [...] *Epistemically*, we do learn about other people's conscious mental states *in part* from their behavior. [..] But *ontologically* speaking, the phenomena in question can exist completely and have all of their essential properties independent of any behavioral output. [Searle 1993: 65-69]

One certainly has to admit that it is *logically* possible - or, as Searle prefers to put it, *ontologically* possible - that a certain behavior exists without the corresponding mental state, or vice versa, that a certain mental state exists without the corresponding behavior. But this statement of a merely *logical* possibility doesn't prove very much. What one would have expected here (and what would be required also in connection with other issues in contemporary philosophy of mind, e.g. the purported existence of Zombies or the so-called "hard problem of consciousness") is rather to show that it is *nomologically* possible that a certain behavior - which is otherwise typical for the presence of a mental state *s* - exists either

without *s* or even without the very possibility of having mental states at all. Searle's core thesis of Biological Naturalism

> (***BN***) Only a living being of a certain complex form will ever have Intentionality,

admits of two different interpretationS. In a weak version it merely expresses the *contingent* empirical fact that so far in nature Intentionality has been observed only as a feature of living beings of a certain complex form. The strong version would amount to the claim that it is nomologically true, i.e., *necessary* according to the laws of nature, that Intentionality occurs only in living beings of a certain complex form. Evidently it is this strong variant of Biological Naturalism which would be required to refute the central premise of the Robot Reply, i.e. the assumption that robots might be able to perceive, to understand language, or to have other forms of mental states, consciousness, and Intentionality.

The thought experiment of the silicon brains fails to provide any compelling reason for this strong form of (***BN***). While the *second* variant of the scenario of the silicon brain was designed to show the logical possibility of a Zombie-like robot's brain *without* consciousness, Searle was forced to admit also, in the *first* scenario, the logical possibility of a non-biological brain which does sustain an ordinary rich mental life. The form of successive presentation of the three variants may have been meant to suggest that these alternatives are *equally* possible. Upon closer analysis, however, one will discover some important differences. To begin with, the third scenario where the person with the silicon brain becomes totally paralyzed while fully retaining her mental life presents a situation which is not only *logically* possible, but also - as was pointed out by Searle himself - *empirically* possible. Such a person strongly resembles a patient with so-called Guillain-Barré syndrome who is "completely paralyzed but also fully conscious". However, such a patient will not at all be diagnosed to be "obviously brain dead"; in this respect Searle's description of the unfortunate situation of the person of the third scenario is incorrect.

Much more disturbing than this incorrectness, however, are certain oddities in the description of the *second* scenario where the person with the silicon brain is assumed to experience a progressive death of her mental life without any change in her (verbal and nonverbal) behavior. The assumption that someone's "conscious experience slowly shrinks to nothing" would entail in particular that she gradually loses all kind of perception. Thus Searle assumes the poor person to be, subjectively, totally blind while her behavior shows all signs of normal visual perception so that, objectively, she can "see" (and is even able to tell what she "sees"). Interestingly, the clash between the subjective and the objective side of the situation is brought to the patient's attention by *hearing* the doctors say that they are holding up a red object in front of her. The diligent reader will wonder, however, how a person who is supposed to have lost *all capacities of perception* may yet be able to *hear* what the doctors say, or *hear* how her own mouth miraculously utters the words 'I see a red object in front of me'. The above quoted description thus contains a slight self-inconsistency which would require a more careful reformulation of the second scenario. For example, one might consider the variant of a person who - while retaining her capacity to hear - is gradually turning "subjectively blind" even though, from the doctor's objective perspective, she is still able to see. This case would be a close analogue of the phenomenon of so-called "blindsight" which is often referred to in recent discussions of the problem of phenomenal consciousness. Yet it seems rather doubtful if, and in which sense, Searle's fiction of a person whose conscious experience shrinks to nothing, while her externally observable behavior remains the same, really is *possible*.

For the sake of completeness let it be pointed out that the *first* scenario illustrating the logical possibility of a silicon brain (and hence of a robot's "brain") with consciousness also contains some rather questionable assumptions. Searle himself expressed his belief that "it is empirically absurd to suppose that we could duplicate the causal powers of neurons entirely in silicon" and he confessed that this belief was just an empirical claim on his part for which he didn't even try to give further substantial reasons. However, such grounds are easily discovered once one gives up to talk only superficially in the lump of the "causal powers of neurons" and considers instead the biological functions of the brain and its neurons in some more detail. Like many other philosophers also Searle seems to work with a grossly oversimplified picture of the human brain according to which its only function consists in transmitting electric impulses. *This* function may perhaps be duplicated by silicon chips. But the single neurons and the entire brain are known to display certain other functions which apparently can never be duplicated by electronic substitutes like silicon chips alone. Just think of the complex *biochemistry* of the brain! A silicon chip will certainly not be able to secrete neurotransmitters or hormones. Hence on the basis of what is currently known about the role of neurotransmitters and hormones for the emergence, e.g., of feelings and emotions, it seems safe to conclude that a person with a silicon brain would never have any *feelings*. And this quite elementary observation appears to have some important further consequences for the main issue of this paper, i.e. for the question whether a robot might be developed which is able to perceive its environment and to understand natural language.

7 Are Conscious Robots Possible?

In a 1994 paper on "The practical requirements for making a conscious robot", D. Dennett describes the outlines of a research project "to design and build a humanoid robot, Cog, whose cognitive talents will include speech, eye-coordinated manipulation of objects, and a host of self-protective, self-regulatory and self-exploring activities":

> One of its most interesting "innate" endowments will be software for visual face recognition. Faces will "pop out" from the background of other objects as items of special interest to Cog. It will further be innately designed to want to keep its mother's face in view, and to work hard to keep mother from turning away. The role of "mother" has not yet been cast [...]. Unlike a human infant [...] there is no reason why Cog can't have a whole team of mothers, each of whom is innately distinguished by Cog as a face to please if possible. Clearly, even if Cog really does have a *Lebenswelt*, it will not be the same as *ours*. [Dennett 1998: 162]

Second, Cog is expected to become able to understand and speak natural language:

> One talent that we have hopes of teaching to Cog is a rudimentary capacity for human language. [...] We are going to try to get Cog to build language the hard way, the way our ancestors must have done, over thousands of generations. Cog has ears [...] and some special-purpose signal-analyzing software is being developed to give Cog a fairly good chance of discriminating human speech sounds [...]. Cog will also have to have speech synthesis hardware and software [...]. It is important to have Cog as well-equipped as possible for rich and natural interactions with human beings [...].
>
> Obviously this will not work unless the team manages somehow to give Cog a *motivational structure* that can be at least dimly recognized, responded to, and

> exploited by naive observers. In short, Cog should be as human as possible in its *wants* and *fears, likes* and *dislikes.* [...] Cog won't work at all unless it has its act together in a daunting number of different regards. It must somehow *delight* in learning, *abhor* error, *strive for* novelty, recognize progress. It must be *vigilant* in some regards, *curious* in others, and deeply *unwilling* to engage in self-destructive activity. [Dennett 1998: 163/4]

But how could a robot ever possess any of the psychological features mentioned in this passage (and set off by me in italics)? How could a robot ever have real feelings, wishes, and needs? Toward the end of his paper Dennett points out to the "deliberate decision by Cog's creators to make Cog as much as possible responsible for its own *welfare*":

> Cog will be equipped with some innate but not at all arbitrary preferences, and hence provided [...] with the concomitant capacity to be "bothered" by the thwarting of those preferences, and "pleased" by the furthering of the ends it was innately designed to seek. Some may want to retort: "This is not *real* pleasure or pain, but merely a simulacrum". Perhaps, but on what grounds will they defend this claim? [Dennett 1998: 169]

Now, it seems to me that there exists at least one decisive reason to infer that a robot produced out of metal, wires, and silicon chips will never be able to *feel* real pleasure and pain (although it may be possible to build Cog in such a way that it *simulates* pain and pleasure *behavior*). The reason being that even the most elementary feelings of hunger and thirst, pain and lust, presuppose the existence of a body which is receptive of biochemical substances and which can react to biochemical agents in a certain specific way. Already the most primitive feelings of pain and lust appear to depend on the secretion and reception of certain biochemical substances which, *for nomological reasons,* can only be produced by appropriate biological tissues but not by wires and silicon chips. Hormones and endorphines will unfold their effects only on the organs of a living being but not on the inorganic components of a robot.

This observation, which of course has to be backed up by further neurobiological investigations, at least partially confirms Searle's position of biological naturalism. Although the foregoing considerations do not prove that by nomological necessity *all* mental states must be confined to the realm of living beings (with higher developed brains), we may ascertain:

> (***BN****) Only a living being of a certain complex form will ever have *feelings*.

And from this it appears to follow that a robot, even if it were constructed so as to be able to perceive its environment and to describe the results of these observations in terms of a natural language, it would still be unable to use language as a tool or medium of *communication*. Pragmatic theories of meaning such as Searle's speech act theory or Grice's theory of conversation and linguistic convention make clear that already the utterance of a simple observational statement such as 'There is a yellow rose standing in front of me' normally presupposes that the speaker, *s*, has certain mental states (or propositional attitudes) which in view of principle (***BN****) a *robot* could never have. For whoever utters such a proposition *p* is not only supposed to *believe* or *know* that *p*; but *s* is also expected to believe that the hearer, *h*, *wants to know* that *p*; furthermore *s* himself is supposed to *want* that *h* understands what *s* has said; *s* normally *wants* that *h* believes that *p*; etc. Communication generally presupposes that both speaker and hearer have certain intrinsic wants and desires, interests and preferences.

And this in turn presupposes that each potential partner of a genuine conversation has certain feelings, or at least the capacity of having feelings.

8 Literature

Block, Ned [1997]: "On a Confusion about a Function of Consciousness", in N. Block, O. Flanagan & G. Güzeldere (eds.), *The Nature of Consciousness*, Cambridge, Mass. (MIT Press), 375-415.

Chalmers, David J. [1995]: "Facing up to the problem of consciousness", in *Journal of Consciousness Studies* 2, 200-219.

Chalmers, David J. [1996]: *The Conscious Mind - In Search of a Fundamental Theory*, New York (Oxford University Press).

Childers, Peter G. & Schank, Roger S. [1984]: *The Cognitive Computer*, Reading, Mass. (Addison Wesley).

Dennett, Daniel C. [1981]: "Three Kinds of Intentional Psychology", in R. Healey (ed.), *Reduction, Time and Reality*, Cambridge (Cambridge University Press), 37-61.

Dennett, Daniel C. [1994]: "The Practical Requirements for Making a Conscious Robot". Reprinted in D. C. Dennett, *Brainchildren – Essays on Designing Minds*, Cambridge, Mass. (MIT Press), 153-170.

Dreyfus, Hubert L. [1993]: *What Computers Still Can't Do*, Cambridge, Mass. (MIT Press).

Grice, H. Paul [1957]: "Meaning", in *The Philosophical Review* 66, 377-388.

Lenzen, Wolfgang [1998]: "Zombies, Zimbos und das »schwierige Problem« des Bewußtseins", in H.-D. Heckmann & F. Esken (Hrg.), *Bewußtsein und Repräsentation*, Paderborn (Schöningh), 255-281.

Linke, Detlef [1990]: "Personalität ohne Gehirn - Medizinische Ethik im kognitivistischen Zeitalter", in *Information Philosophie* 5, 5-15.

Meggle, Georg [1981]: *Grundbegriffe der Kommunikation*, Berlin (de Gruyter).

Nagel, Alexandra H. M. [1997]: "Are Plants Conscious?", in *Journal of Consciousness Studies* 4, 215-230.

Rheinwald, Rosemarie [1992]: "Das Chinesische Zimmer als Test des Turing-Tests?", in *Philosophische Rundschau* 39, 133-156.

Searle , John [1969]: *Speech Acts*, Cambridge (Cambridge University Press).

Searle, John [1980]: "Minds, Brains and Programs", in *The Behavioral and Brain Sciences* 3, 417-457; here quoted according to the reprint in J. Haugeland (ed.), *Mind Design*, Cambridge, Mass. (MIT Press), 282-306.

Searle, John [1983]: *Intentionality: An Essay in the Philosophy of Mind*, Cambridge (Cambridge University Press).

Searle, John [1984]: "Intentionality and Its Place in Nature", in *Synthese* 61, 3-16.

Searle, John [1984a]: *Minds, Brains, and Science* ("The 1984 Reith Lectures"), publ. by the British Broadcasting Corporation.

Searle, John [1990]: "Is the Brain's Mind a Computer Program?", in *Scientific American* 262 (January), 20-25.

Searle, John [1992]: *The Rediscovery of the Mind*, Cambridge, Mass. (MIT Press).

Turing, Alan M. [1950]: "Computing Machinery and Intelligence", in *Mind* 59, 433-460.