Mario Radovan
FET Pula, Sveučilište Rijeka
E-mail: mradov@efpu.hr

UDK: 007.52
Original Scientific Paper

Mind and Computation (II)

The paper deals with the fundamental problems of cognitive science, starting from the epistemic and ontological limitations which are immanent to the very attempt to describe mental phenomena in terms of objective (scientific) language, up to the problems of the formal representation of common-sense knowledge. A many-level model of the cognitive system has been proposed; in that context, we analyse the Classical and Connectionist approach to the description of the cognitive system, and we argue that (1) these two approaches should be conceived as two different levels of speech about the same phenomena, and that (2) they face essentially the same basic problems. The second part of the paper discusses various positions concerning the range and limits of artificial intelligence; in that context we put forward the Background and Care hypotheses, both of which call into question the very possibility of the existence of machines with (any) real cognitive abilities. The paper concludes that the requirements which are put before AI should be more realistic (than they usually are) if we are to deal with reasonable research projects.

Key words: mind, cognitive models, intelligent systems, language of thought, computation, connectionism, artificial neural networks, background hypothesis, care hypothesis.

1. Introduction

With the present paper we intend to complete and conclude the critical analysis of the basic ideas and problems in the field of cognitive sciences, with which we dealt in [25]. The relation between *mind* and *computation* has been a subject of treatises for centuries (cf. [17] and [6]), but the development of *computing machines* has dramatically increased its importance. Research concerned with human cognitive abilities and with the possibilities of their artificial replication forms the core of *cognitive science*, which aims to integrate results from various fields such as the *theory of computation*, *artificial intelligence*, *formal linguistics* and *cognitive psychology*.

In the last few decades much has been done in the field of cognitive science; however, it seems that the fundamental problems concerned with the human mind resist all attempts to resolve them by means of computational taxonomy and technology. Indeed, as a kind of reply to Descartes, we could say that the human *knows that he is*, but that he (still) doesn't know *what* he is. Namely, in almost every treatise concerning the human mind, we encounter the claim that the *conscious mind* is a *mystery*. Clark says that it is a "mystery how conscious content is possible at all" [5, p.

224], and Dennett describes "mystery" as "a phenomenon that people don't know how to think about it" [8, p. 21]. In keeping with such a position, Dennett concludes his book *Consciousness Explained* with the confession that his explanation was "far from complete"; namely, he has not proposed a new scientific theory but only a new metaphor [8, p. 455]. The concepts of *consciousness*, *mind* and *self* are, in fact, usually not even defined literally, but only figuratively; moreover, there are claims that it is not possible to give "noncircular verbal definitions" of these concepts, and that their meanings can be best expressed by means of examples [26, p. 83]. Let us see of what the difficulties with the mental consist.

1.1 The observer/observed gap

Science assumes that reality is *objective* in the sense that neither its existence nor its structure depend on a particular observer; to explain a phenomenon scientifically means to describe it from the *neutral* (third-person) point of view. However, there are claims that conscious mental states cannot be described in the neutral (non-personal) fashion because every such state is essentially a *personal/subjective* state. "Mental states are always *somebody's* mental states", says Searle [26, p. 20]; Nagel holds the same position when he claims: "The subjectivity of consciousness is an irreducible feature of reality ... and it must occupy as fundamental a place in any credible world view as matter, energy, space, time and numbers" [22, pp. 7-8]. In other words, both authors hold that "the ontology of the mental is an irreducibly first-person ontology" [26, p. 95]. However, Flanagan holds that "the gap between the subjective and the objective is an epistemic gap, not an ontological gap" [11, p. 221]. Let us first see how the epistemic gap comes about, and why it cannot be eliminated.

The spectator always remains out of the range of his own view; this epistemic necessity perplexed thinkers from Hume and his attempt to see the *proper self*, up to Dennett's attempt to eliminate the epistemic gap by means of the *computational metaphor* of the mind. In the often quoted passage, Hume says: "when I enter most intimately into what I call *myself*, I always stumble on some particular perception or other, of heat or cold, ..., pain or pleasure. I can never catch *myself* ... without a perception, and can never observe anything but the perception" [19, p. 252]. From that, Hume concluded that the self was nothing but a bundle or collection of different perceptions which succeeded each other in perpetual flux. What Hume was searching for, and what he was not able to find, was *the searcher*; but that should not have astonished him because the act of observing necessarily includes the existence of the two sides: *subject* (the observer) and *object* (the observed); and the observer cannot become an object of his own observation because he cannot pass on that side of the epistemic gap which is *observed* (thought). In other words, observed/thought phenomena *became* phenomena "only through the objectifying activity of a subject which transcends objectification in the same sense" [7, p. 43].

Following the computational model of the mind, Dennett defines consciousness as a *virtual machine*. In computer systems the term 'virtual' usually denotes indirect addressing and dynamic resource allocation; in practical terms, that primarily means

that the user does not have direct control (neither has to care) over the allocation of the hardware resources on which his program runs. Dennett sees the danger which is inherent in his attempt, and feels compelled to ask: "If consciousness is a virtual machine, who is the user for whom the user illusion works? I grant that it looks suspiciously as if we are drifting inexorably back to an internal Cartesian Self, sitting at the cortical workstation"; Dennett hopes that "there are ... some ways of escaping that dreadful denouement" [8, pp. 219-20]. But there are not, because there is no sensible way to speak of a *machine* without the *user*. Or, to state the same thing in other words: "The homunculus fallacy is endemic to computational models of cognition and cannot be removed by the standard recursive decomposition arguments" [26, p. 226].

Arguing against the alleged explanatory power of the computational metaphor, Searle rightly concludes that if we suppose that the brain is a computer, we are still faced with the question 'And who is the user?' [26, p. 214]; this is, in fact, just another way to state the homunculus fallacy. However, the same argument can be used to show that the homunculus fallacy is endemic to our *basic cognitive situation* in general. Namely, by paraphrasing Searle's words, we could ask: If we suppose that the brain is *brain*, we are still faced with the question 'And who is the user?'. This question once again points to the fact that it is not possible to observe/think without dividing the *Existent* into two disjunctive parts: the observed/thought and the observer/thinker, or simply, into the object and the subject. In that context, I consider Flanagan's position correct, because even if the *Existent* is ontologically monolithic, nothing can be seen/thought without "introducing" an epistemic gap: and the subject always stands on the "wrong" side of the gap; that could be one of the reasons for the "mysterious nature" of the conscious mind.

1.2 The mental/physical gap

Attempts are made to eliminate the subject/object gap by reducing the mental state to the physical. Such attempts follow the common scientific practice to define the "surface phenomena" in "more basic" (physical) terms. The idea that the mental state can be reduced to the physical assumes that to redefine a given phenomenon in terms of some "lower level" language means also to challenge its "reality" as a phenomenon at some "higher level". But where the mental state is concerned, such an attitude seems to be wrong because no redefinition of my pain or desire in chemical or physical terms can make it less real. A scientific description of a mental state can give its *how,* but not its *what*, which is intrinsically subjective, and can be completely known only to the conscious subject to whom it belongs. In fact, the language of physics "appears to have some limits, and it reaches them at the subjective character of the contents of consciousness" [4, p. 196]. That does not mean that the phenomenon of conscious mental states is nonphysical in nature, but only that the scientific taxonomy cannot express *"what they are like* from the unique perspective of the creature that has them". In other words, the difference is not supposed to lie in the *character* of the thing known, but in the *manner* of the knowing [4, 196]. In essence,

we have no clear idea how to deal with *subjective experience* in an *objectively described world*, so that we are prone to speak of mental states in ways which can often be qualified as banal or incoherent, or both. In that context, many spectacular claims concerning the relation between the human mind and (future) computers could be qualified as rhetoric figures rather than as clear and well-grounded scientific positions.

2. Levels of description

To describe a phenomenon, one needs a system of concepts. If *theory* can be said to be "the conceptual vehicle with which we ... come to grips with the world" [3, p. 117], then *metaphor* could be said to be a conceptual vehicle with which we try to come to grips with the (actually) inexpressible: such attempts of metaphorical speech are intended as a first move towards a scientific theory. For the description of the human cognitive system, the taxonomy of digital computers - understood in the figurative sense - has been taken as a promising starting model.

A computer system can be described at many different *levels of abstraction*; following Winograd and Flores [29], we introduce five levels of description; for each of the levels, we propose an analogous level of description of the human cognitive system.

Physical level - At this level, the computer is seen as a set of elements which operate in accordance with the laws of physics. There are no symbols or operations on this level: at best, we could speak here of *signals* described in terms of the laws of physics. In the human cognitive system, that would be the level of *neuroanatomy* which deals with the material and structural aspects of the brain's neural system.

Logical level - At this level, the system is seen as a network of *logical gates* (the standard "and", "or", "not" gates); here, the system can be described by some binary language. In the human cognitive system, that would be the level of *neurophysiology*, where the brain is represented as a set of *networks of functionally described neurons.*

Representation level - The level of the *symbolic machine language* (assembler); at this level, strings of binary symbols are interpreted as representations of *data* and *operators/commands*. Concerning the human cognitive system, this is the most controversial level of the model; to follow the computer model, we must assume the existence of some kind of "assembly language" in the human brain. The best known proposal of such a language is Fodor's *Language of Thought* [12]; we deal with this in section (3.1).

Communication level - The level of *programming/query languages* by means of which the user exchanges data and instructions/queries with the system. With humans, that would be the level of natural language communication and reasoning; in keeping with the dominant terminology we shall often call it *linguistic level.*

Situation level - The level at which an activity of the computer system is interpreted as *solving a problem*. In the human cognitive system, that would be the level of *understanding* and of *goal-directed activity*.

These levels are defined in functional terms, so that the system can be studied (and, in principle, modelled) on each level of abstraction independently of other levels. Functionally defined systems are *medium independent*; consequently, if the proposed model allowed the complete description of the human cognitive system, and if all the functions from that description were realisable by artificial means, it would follow that the human cognitive system could be replicated in ways and media which are structurally and materially different from the human brain. In the second part of the paper, we claim that the both, the model as well as the perspectives for its realisation, are rather critical. But let us first discuss the two most important levels of the model, the *representation level* (also called the basic *software level*) and the *logical level* (also called the *hardware* level).

3. The classical approach

The Classical approach to cognition (also called the *Symbolic Information Processing (SIP)* approach) aims to describe human cognitive abilities on the basic software level (i.e. on the *representation* or the *machine/assembly language* level). Namely, it is supposed that if we could describe and replicate the human cognitive system on that level, it should also be relatively easy to replicate the features of two higher levels, i.e. to obtain an artificial system that can be said really to *communicate* and to *understand*.

3.1 The LOT hypothesis

We assume that cognitive abilities do not depend on any particular natural language since the same thoughts can be (or *could* be, by coherent extensions) expressed in different natural languages. On the other hand, the linguistic abilities of speakers of different natural languages show the same *structural* properties. These two premises lead to the idea that the human cognitive system contains an *internal language* (innate and common to all humans) in terms of which the cognitive processes take place. This internal language has been called the *Language of Thought (LOT)* or *Mentalese*. To find out what the hypothetical LOT level of the cognitive system looks like, we proceed by the following line of thought: (1) human linguistic abilities are characterised by certain structural properties; (2) sentences of natural language express/mirror thoughts; (3) therefore, the cognitive abilities should have the same structural properties as the linguistic one; (4) LOT is that formal system which offers the *best explanation* of these structural properties.

According to Fodor and Pylyshyn [14], the basic structural properties of the human linguistic system are: *productivity*, *systematicity*, *compositionality*, and *syntax/semantics coherence*. These properties can be best explained by assuming a

representational and *combinatorial* nature of the linguistic system. The concept of "representational nature" implies that language units represent entities (in the world), while "combinatorial nature" implies that the validity/meaning of *complex* linguistic units is defined in terms of the validity/meaning of their constituent parts. In accordance with the starting position concerning the relation between linguistic and cognitive abilities, it has been assumed that at some *thought-producing level* the human cognitive system could be defined as a representational and combinatorial system: therefore, that it could be described by a kind of *representation language of combinatorial syntax and semantics.* LOT is taken to be that language.

The LOT hypothesis further holds that: (1) the states of some "points" of the brain form the representation of some proposition, and with it a representation of some state in the world; (2) propositions are *composite* entities, and the same holds for their mental representations: they are composed of *mental atoms*, the minimal content-bearers; (3) tokens (*physical items/signs* in the brain) which record the *same content* are of the *same form/type*: that form is the *mental symbol* of that specific semantic content; (4) there is a coherent relation between the *syntax/form* level and the *semantic/content* level of the operations which take place in LOT seen as a formal system. The LOT hypothesis does not say anything about the exact forms and contents of mental atoms; but starting from the assumed *existence* of such minimal representation items of fixed syntax and semantics, it offers an explanation of the reasoning processes. In a nutshell, the LOT hypothesis claims that the syntactic properties of a representation item can be reduced to its *shape*, as a *physical* property. This further means that *causal interactions* of tokens (which depend on their physical properties) are determined by syntactic properties of the *mental symbols* which they token. Consequently, although the physical properties "onto which the structure of the symbols is mapped" are those that "cause the system to behave as it does" [14, p. 14], the human cognitive system can be conceived as an *automated symbol system.* In other words, the cognitive processes can be equally seen as *causal sequences* of tokenings (of mental symbols) and as a *formal (rule-driven) symbol manipulation.*

LOT is the core element of the Classical/SIP model of cognition. To complete the model, Fodor also introduced a set of functional "boxes" which can be described as special-purpose processors. For example, to *believe* P would mean to have a token of the mental symbol 'P' in the *belief box*, and to *hope* P would mean to have a token of the *same* mental symbol in the *hope box*, and so on, "a box for every attitude that you can bear toward a proposition" [13, p. 17]. And for actions, there is an *intention box*: when you intend to make it true that P (i.e. *do/make* what 'P' says), you put into the intention box a token of the mental symbol for P; "the outcome is that you then behave in a way that (ceteris paribus) makes it true that P" [13, p. 136].

3.2 Comments on the SIP/LOT

LOT is a hypothesis, and it should be evaluated on the grounds of its explanatory power and its pragmatic effects. Many hold that the SIP model (based on the LOT hypothesis) is the best theory of cognition we have; Fodor claims that "the cost of not

having a Language of Thought is not having a theory of thinking" [13, p. 147]. However, there are also opponents of the computational approach to cognition in general, and of the LOT hypothesis in particular. Searle, for example, says: "The brain produces the conscious states that are occurring in you and me right now. But ... that is the end of the story. There are brute, blind, neurophysiological processes and there is consciousness, but there is ... no mental information processing, ..., no language of thought, and no universal grammar" [26, pp. 228-9]. Searle's position doesn't seem coherent to me. Namely, *every* theory/model implicitly imposes some structure onto reality; LOT does this, but so does neurophysiology (cf. [9]). It could be a psychological fact that by approaching the level of the physical we *feel* as if we are approaching Truth/Reality, but that is only an illusion: science is a pragmatic enterprise, and *all* theories are in essence only hypotheses.

The main weakness of the LOT hypothesis (and of the SIP) taken as an *explanation* of human cognitive abilities, concerns the semantics of mental atoms. A LOT-based system is a model of the internal level of the human cognitive system. Now, in order to parallel the *syntax/semantic coherence* of the outer (linguistic) level, the LOT system cannot be merely a *syntactic* engine, because such systems can generate only new meaningless marks from the existing ones. To pass from marks to *thoughts/meanings*, the LOT hypothesis must hold that representation items have not only fixed forms, but also *innate fixed meanings*. A justification for such an assumption could run like this: (1) if thoughts have meanings, then meanings must come from somewhere; (2) let us take it that they come from the semantic properties of the *basic mental items*. Within Fodor's model, such an argument could suffice, but it makes the model rather speculative: it is still useful as a working hypothesis in AI, but of limited value as an *explanation* of human cognitive abilities. However, let us note that adherents of the SIP approach are not, in fact, even supposed to answer *what makes* the atomic items have semantic properties. The SIP approach is defined in *functional* terms: it assumes the existence of a fixed set of "basic building blocks" out of which all other items and *explanations* are constructed, but which are *by definition* cognitively impenetrable. In other words, by having no way to eliminate the perennial gap between the *objective* and the *subjective*, the SIP/LOT approach simply bridges it by an assumption which "works" inside the model, but nobody knows how.

4. The connectionist approach

While the Classical approach starts from human linguistic abilities, Connectionists aim to develop systems with cognitive abilities starting from the *hardware* level, modelled as a kind of imitation of the brain's neural structures. Some adherents of this approach claim that "the neurocomputational alternative promises to provide some solutions where the older [Classical] view provided only problems" [4, p. 252]. That could be partially true; however, in sections (5) and (6) we claim that the Connectionist approach does not have solutions for the most essential problems

concerning machine intelligence/understanding, which seems to be common to both approaches.

4.1 Artificial neural networks

An *artificial neural network (ANN)* consists of a set of interconnected units (nodes), divided into input units, output units, and hidden units which mediate the spread of activation between input and output units. A node is characterised by a *variable* representing its *level of activation* and by a *constant* representing its *threshold*: when the input activation of a node exceeds its threshold, activation propagates to the other nodes with which that node is connected. Links are *weighted*, with weights determining the relative quantity of activation they may carry. An input to the network is a *vector of signals* clamped on input nodes (to every node a value). Triggered by the input, activations spread throughout the network in a way determined by the input pattern, node thresholds, links, and weights of the links. The output of the system is the vector formed of the activation values of the output nodes when the network settles down into a steady state. ANNs are also called *vector transformation systems*.

An ANN acquires knowledge by *being trained* on a set of examples. The system usually starts with a random distribution of unit thresholds and weights; after every exposure to a training exemplar, states of the output units are compared with the *desired* output pattern, and the weights/thresholds of the units are gradually changed until the output pattern (for the given input) becomes equal to the desired one. The same process is repeated with each training exemplar. Adjustments made during training with one exemplar may distort the knowledge acquired through former training exemplars, so that the training process must be cyclically repeated until the network reaches a configuration which correctly transforms *all* the exemplars from the training set. It is said that the training process "extracts the *statistical central tendency*" of the training exemplars, forming with this a "*prototype-style* knowledge representation" of the characteristic features of the exemplars [5, p. 20]. An ANN shows its knowledge (acquired by training) when it is requested to process *new* inputs of the same type as those with which it was trained. Let us mention that ANNs are characterised by *holistic* knowledge storing, in the sense that any node can take part in the encoding of any piece of knowledge contained in the network. It means that in ANNs there are no context-independent data records which could be said to represent natural language semantic units: any node can take part in *encoding* many things, but it *represents* no particular thing.

4.2 Comments on ANNs

Skills such as riding a bicycle or playing the piano (and many others) do not consist in verbal knowledge, and exercising them does not require explicit rule-driven reasoning; moreover, we usually cannot even describe them in a verbal/symbolic form. Humans acquire such skills by *practice*: consequently, some kind of *training-*

based cognitive model should lead to the understanding and also to the replication of the systems which posses such skills. This is the strongest argument for the Connectionist approach. On the other hand, the list of opened problems is rather long; the most important among them concern the *representation* of input/output data (because not all data can be easily stated in vector form), and the *learning algorithm* which could successfully decide *which* weights and thresholds should be changed (during the learning process) and *how* they should be changed (i.e. in what direction and to what degree). Without satisfactory solutions to these two problems, every excessive enthusiasm for the Connectionist approach seems completely ungrounded.

The connectionist approach has an inherent drawback concerning the *explanation* of its own results. Namely, by renouncing symbols and rules, Connectionism has deprived itself of means by which it could form any scientific explanations, so that it must borrow some "alien" explanatory means to qualify as a *scientific* activity. To overcome this drawback, various techniques such as *cluster analysis*, by means of which one can generate a "static symbolic description of a network's knowledge" have been used [5, p. 33]. Such descriptions do not mean that the symbols (which they use) exist as syntactic items in the network: they are only *post-hoc* semantic explications of what the network *knows/does*, and not of what is going on inside the network. However, this drawback could be interpreted also as an advantage; namely, as Clark put it, concerning the explanation, the Connectionist approach is "both sound and problematic": it is sound because it avoids projecting a coarse symbolic language onto the cognitive mechanism itself; it is problematic because, by the same token, it deprives itself of explanatory means [5, p. 67].

It is often argued that the Connectionist approach should confine itself to the problems of the *physical implementation* of the cognitive system defined in the SIP fashion. In that case, Connectionism would not be a cognitive theory, but only an implementational model for the SIP theory of cognition. However, although advocates of the SIP approach, Fodor and Pylyshyn point to cases where Connectionism by itself offers the most suitable approach; for example, they hold that "the input to the most peripheral stages of vision and motor control *must* be specified in terms of anatomically projected patterns", and hence, that "at these stages it is reasonable to expect an anatomically distributed structure to be reflected by a distributed functional architecture" [14, p. 63]. In other words, independently of its qualities as the model of the hardware implementation of the SIP cognitive model, Connectionism seems to be the right approach for the development of *input/output units* within a global SIP-oriented model of the human cognitive system.

Finally, there are also rather curious arguments in the favour of the Connectionist approach. For example, Churchland says: "In humans, ... the basic unit of cognition is the *activation vector* ... the basic unit of computation is the *vector-to-vector transformation* ... the basic unit of memory is the *synaptic weight configuration*. None of these things have anything essential to do with sentences ... or with inferential relations between them" [4, pp. 322-3]. It seems that Churchland is not aware how much his words resemble the micro-level description of the classical digital computer,

where computation is nothing else but a vector-to-vector transformation. (In classical machines, vectors consist only of the binary values, but that is neither necessary nor essential.) However, it would be extremely difficult (but not impossible!) to *write large programs* in a vector-to-vector transformation fashion: that was why high-level programming languages together with compilers were developed. But a program written in a high-level language still defines *the same phenomena* (i.e. data and processes) as the vector-to-vector transformation description does. The same should be the case with the human cognitive system: hence, there is not much sense in claiming that any of the levels of description has any *a priori* advantage, or that the entities from one level of description (i.e. "vectors") have not "anything essential to do" with the entities of another level of description (i.e. with "sentences").

5. Mind and the artificial

Let us now see a few typical positions concerning *Artificial Intelligence (AI)* in the context of what has been said about the phenomenon of the mental and about approaches to the human cognitive system. Speaking of the "very idea" of AI, Haugeland says: "The fundamental goal [of AI] ... is not merely to mimic intelligence or produce some clever fake. ... AI wants only the genuine article: *machines with minds*, in the full and literal sense. ... Namely, we are, at root, *computers ourselves*" [17, p. 2]. On the other hand, Searle declares that he intends "to put a final nail in the coffin of the theory that the mind is a computer program" [26, p. xi]. Searle is right in his insisting that the human mind is not intrinsically a computer program; however, he completely neglects the possible pragmatic value of the *computational interpretation* of the human mind in the context of our efforts to develop useful model of the human cognitive system as well as to develop more efficient computers. But in the last instance, I consider all such claims primarily as rhetoric figures, because by speaking in an imprecise manner, everything can be *interpreted* as a machine and nothing can be *proved* to be intrinsically a machine, not even computers themselves if seen on the "wrong level" of description.

Concerning the question of the *intelligence* of artificial systems, we face the same problems. It has become fashionable to claim that according to the "new approach" to AI, "to design an intelligent system, one has to give it all the properties of intelligent creatures", including "intentionality, consciousness and autonomy along with ... adaptivity" [15, p. 483]. Gams recognizes that to obtain all that "will be much more difficult than previously expected" [15, p. 488]. However, I don't know who "previously expected" that to construct an artificial system with the above stated properties could be *less* difficult than anything else! Moreover, such requirements concerning machine intelligence (which include consciousness and intentionality) put AI in a rather curious position; namely, they make it virtually impossible to construct anything that would be at the same time "intelligent" and "a machine"! Such requirements are simply so strong that they render it impossible even to try to undertake any reasonable step toward something that could be accepted as a "first approximation" to an intelligent artificial system. Faced with such a situation, we are

constrained to abandon the very idea of constructing intelligent systems, or to change our criteria about the intelligence of artificial systems. In this context, I hold that as long as we do not know how we could construct a *conscious machine* (and we could hardly know that as long as we do not know even how natural consciousness comes about!), we should define the intelligence of an artificial system in strictly *behavioural fashion*. In other words, if we insist that a chess program which can beat a chess grandmaster is nevertheless not intelligent, then not only do we not actually have intelligent systems but we also do not have *any idea* how we could construct one.

The criticisms of the results and perspectives of AI concern primarily the Classical approach to AI. That approach follows the tradition of Western thought which holds that to understand/construct something we must first have a *theory*. In that context, to produce an artificial intelligent system, one is supposed to find out and describe (at some suitable level of description) the basic elements and laws of the *natural* intelligent (sub)system, and represent that knowledge in some formal language which is implementable on the computer. The first attempts in this direction were limited to the selected *micro-worlds* (i.e. to selected sets of abilities/knowledge); but it turned out that the intelligence of such systems, without large amounts of *common-sense knowledge*, was radically limited. On the other hand, all the proposed methods for the formal representation of common-sense knowledge turned out to be problematic and highly controversial. The most promising attempt in that direction, the *CYC project*, was started by Lenat in 1984, with the aim of building a system whose knowledge would contain most of what humans call common-sense knowledge. ("CYC" is an abbreviation of en*cyc*lopedia; after some initial amount of explicitly inserted knowledge, the system was supposed to continue to learn automatically from media such as books, newspaper, etc.) The CYC project follows the SIP approach; it uses explicit knowledge representation (in a formal language of sentential form) and an extensive formal inference system (with heuristics). It was thought that such a system could serve as the formalized common-sense background for expert systems of various kinds. Although the project is not fully completed, it has been claimed that CYC is not going to attain its mark (or that it has already missed it); for example, Mickie holds that CYC has not attained "even the semblance of human-level knowledge and intelligence" [21, p. 464].

On the other hand, the Connectionist approach departs from the theory-oriented tradition, and attempts to replicate intelligent behaviour without its explicit formal description (i.e. without a theory). There are claims that ANNs will avoid the main drawbacks of the SIP systems; however, it is too early to judge the possibilities of that approach, because the existing ANN systems are still very limited, so that "the same common-sense knowledge problem, which has blocked the progress of symbolic representation techniques for fifteen years, may be looming on the neural net horizon" [10, pp. 438-39]. Finally, with ANN we encounter the same basic problem as with SIP: namely, if an ANN system is to reach the *real* (human-like) intelligence, it is supposed that such a system should also "share our [human] needs, desires, and emotions and have a human-like body with the same physical movements, abilities

and possible injures" [10, p. 440]; let us add here that it should also be conscious of *its own approaching death*. In short, with ANNs we arrive at the same excessively strong requirements: too strong to permit us even to try to do anything in the direction of their realisation. Hence, one of the basic tasks of artificial intelligence should be to define in a more precise and *more realistic* way the proper goals. Without that, a great part of the discussions will continue to belong to one of the three traditional classes: vague speculations, unrealistic promises, and lamentations of failed expectations.

6. Cognition and computation

Our analysis has been concerned primarily with the *methodological* and *epistemic* differences between the Classical and Connectionist approaches. But is there any deeper, *ontological*, difference between these two approaches? It seems as it should be, since they look so different, but it is hard to say of what the difference consists. Namely, every ANN can (in principle) be simulated on a SIP system, and every SIP system can (in principle) be simulated by a set of specialised ANNs (e.g. every basic function of the SIP system is simulated by an ANN). Therefore, the two cognitive models have *the same expressive power*. Indeed, we have already argued that they describe the same phenomenon on two different levels, and that they can be generally conceived of as the "hardware" and the "software" descriptions of the same system.

Concerning the relation between cognition and computation, there are claims that classical computers, as symbol processing systems, are inherently unable to acquire any real cognitive ability. A symbol system can only manipulate that knowledge which can be expressed in some symbolic language; and it has been argued that human knowledge cannot be stated in purely linguistic form. Winograd and Flores hold that the two essentials of the human cognitive situation are: (1) man is always already situated in some *cognitive background* which cannot be explicated (and hence not formalised); and (2) knowledge consists in *concernful acting* ("care", in Heideggerian terms) and not in mere information possessing (neither of the symbolic/SIP nor of the neural/ANN kind). According to the *Background* hypothesis, the meaning of an expression is the result of its *interpretation* against some given background; the unspoken (i.e. the background) determines the meaning more than what has been said explicitly. "Every explicit representation of knowledge", says Winograd, "bears within it a background of cultural orientation that does not appear as explicit claims, but is manifest in the very terms in which the 'facts' are expressed and in the judgment of what constitutes a fact" [28, p. 453]. An attempt to *explicate* all the content of the background would be not only endless but also useless because assertions without any background have *no meanings* at all (cf. [26]). Consequently, the category of meaning is not applicable inside formal systems because symbol manipulation, by itself, without an *outer* interpreter, is simply senseless (and so is ANN vector transformation). On the other hand, according to the *Care* hypothesis, human communication is a form of *concernful social action*, and not mere transmission of information. Social action implies

commitment: a normal human being cannot be said to *understand* an assertion (heard or said) without being somehow committed to its content. And to be committed, one must be *somebody* (must be an *I*), which an automated system is not: hence, automated systems cannot really understand. (Such claims directly challenge the Classical approach, but since we do not see any ontological differences between SIP and ANNs, we assume that they also equally challenge the Connectionist approach.)

On the basis of the *Background* and *Care* hypotheses, Winograd and Flores claim that computers cannot even *in principle* acquire any real cognitive abilities. And that also means that human cognitive abilities can be neither explained nor replicated by computing systems (of any kind), because human cognition cannot be reduced to mere computing. Let us note that this line of reasoning implicitly rejects the independence of *cognition* from the *subjective mental state*, the independence which has been one of the basic assumptions in cognitive science. I hold that Winograd and Flores are right, and that a *Care-less* automated system can neither *communicate* nor *understand* in the sense in which people do. And what about *thinking* and *intelligence* without *understanding*? With that, we have returned to the starting problem of how to evaluate (or speak of) the intelligence of artificial systems. As already stated, we hold that the intelligence of such systems should be judged in a *behavioural fashion*, despite all the drawbacks of such a decision. Turing [27] was the first to put forward the question of machine thinking/intelligence; his paper begins with the words: "I propose to consider the question 'Can machines think?'". He also proposed a *test* (later known as the *Turing test*) on the basis of which a given machine could be qualified as intelligent. The turing test is highly controversial and much criticised; e.g. Hofstadter says that "the Turing test seems to settle dogmatically on some form of behaviourism, or (worse) operationalism, or (worse still) verificationism" [18, p. 93]. That may be true; but it is equally true that we do not know how we could effectively/reasonably approach the problem in some other way. Epistemic and methodological changes which bring about the passage from SIP to ANNs are not the decisive step towards *true* (i.e. human-like) understanding and intelligence. And we do not actually have any idea what *the decisive* step in that direction should look like. In that context, claims that humans are machines, or that to become intelligent, machines should share our needs, desire, and emotions, do not tell us as much (especially not in the operational sense) as it might appear at first sight.

7. Conclusion

This paper has put forward and discussed some of the fundamental problems from the wider scope of cognitive science. In that context, we have argued that: (1) an attempt to understand the conscious mind must face the gap between the ontologically monolithic picture of the Existent and the inherent duality of the epistemic act of observing; (2) the actual objective language of science has its inherent limitations, which become evident when we try to represent subjective states in terms of objective language; (3) the Classical and the Connectionist approaches, which are claimed to be radically different, are primarily two different levels of

description of the same phenomenon, and hence they face the same basic problems; (4) we do not know how we could create an artificial system that could "really understand" without being conscious (and *personally involved*) in the social world, but at the same time we are not able to conceive how an artificial system *could* have such features; (5) finally, we must tone down the requirements and expectations of AI if we are to deal with realistic projects and reasonable discourse in that scope.

8. References

[1] Armstrong, D. M.: *A Materialist Theory of the Mind*, Routledge, 1995.

[2] Bojadžiev, D.: Godel's Theorems for Minds and Computers, *Informatica*, Vol. 19 (1995), pp. 627-634.

[3] Churchland, M. P.: *A Neurocomputational Perspective: The Nature of Mind and the Structure of Science*, The MIT Press, 1992.

[4] Churchland, M. P.: *The Engine of Reason, the Seat of the Soul*, The MIT Press, 1995.

[5] Clark, A.: *Associative Engines: Connectionism, Concepts, and Representational Change*, The MIT Press, 1993.

[6] Copeland, J.: *Artificial Intelligence: A Philosophical Introduction*, Blackwell, 1993.

[7] Copleston, F.: *A History of Philosophy, Vol. VII*, Search Press / Paulist Press, 1963.

[8] Dennett, D. C.: *Consciousness Explained*, Penguin Books, 1993.

[9] Devitt, D., Starenly, K.: *Language and Reality*, Blackwell, 1987.

[10] Dreyfus, L. H., Dreyfus, E. S.: 'Making a Mind vs. Modeling the Brain: AI Back at a Branchpoint', *Informatica*, Vol. 19 (1995), pp. 425-441.

[11] Flanagan, O.: *Consciousness Reconsidered*, The MIT Press, 1992.

[12] Fodor, J. A.: *Language of thought*, Harvard University Press, 1975.

[13] Fodor, J. A.: *Psychosemantics*, The MIT Press, 1987.

[14] Fodor, J. A., Pylyshyn, Z. W: 'Connectionism and cognitive architecture: critical analysis', *Cognition*, Vol. 28 (1988), pp. 3-71.

[15] Gams, M.: 'Strong vs. Weak AI', *Informatica*, Vol. 19 (1995), pp. 479-493.

[16] Gillies, D.: *Philosophy of Science in the Twentieth Century*, Blackwell, 1993.

[17] Haugeland, J.: *Artificial Intelligence: The Very Idea*, The MIT Press, 1986.

[18] Hofstadter R. D., Dennett, C. D.: *The Mind's I*, Penguin Books, 1981.

[19] Hume, D.: *A Treatise of Human Nature*, Oxford University Press, 1951.

[20] Kenny, A.: *The Metaphysic of Mind*, Oxford University Press, 1989.

[21] Michie, D.: '"Strong AI": an Adolescent Disorder', *Informatica*, Vol. 19 (1995), pp. 461-468.

[22] Nagel, T.: *The View from Nowhere*, Oxford University Press, 1986.

[23] Ortony, A. (ed): *Metaphor and Thought*, Cambridge University Press, 1993.

[24] Radovan, M.: 'On the Computational Model of the Mind', *Informatica*, Vol. 19 (1995), pp. 635-645.

[25] Radovan, M.: 'Mind and Computation', *Zbornik*, FOI Varaždin, Vol. 19 (1995), pp. 47-67.
[26] Searle, J.: *The Rediscovery of the Mind*, The MIT Press, 1992.
[27] Turing, A.: Computing Machinery and Intelligence, *Mind*, 59 (1950); reprinted in Hofstadter and Dennett (1981), pp. 53-67.
[28] Winograd, T.: 'Thinking Machines: Can There be? Are We?', *Informatica*, Vol. 19 (1995), pp. 443-459.
[29] Winograd, T., Flores, F.: *Understanding Computers and Cognition*, Addison-Wesley, 1987.

Received: 1996-05-21

Radovan M. Um i računanje (II)

Sažetak

Članak se bavi temeljnim problemima kognitivne znanosti, počevši od epistemoloških i ontoloških ograničenja, koja su imanentna samom pokušaju opisa mentalnih fenomena u terminima objektivnog (znanstvenog) jezika, pa do problematike formalnog zapisa općih (zdravorazumskih) znanja. Predložen je jedan višerazinski model kognitivnog sustava; u tom kontekstu, analizirani su klasični i konekcionistički pristup opisivanju kognitivnih sustava, uz zaključak da: (1) ta dva pristupa treba promatrati kao dvije različite razine govora o jednom te istom fenomenu, te (2) da se ta dva pristupa suočavaju s istim temeljnim problemima. Drugi dio članka razmatra stavove o dosezima i ograničenjima koja su svojstvena umjetnoj inteligenciji; u tom kontekstu iznosimo hipotezu Pozadine (Podloge) i hipotezu Brige, koje zajedno dovode u pitanje samu mogućnost postojanja strojeva s (bilo kakvim) stvarnim kognitivnim sposobnostima. Članak zaključuje da očekivanja koja se postavljaju pred umjetnu inteligenciju trebaju biti znatno realističkija (nego što to obično jesu) ukoliko se u tom području želimo baviti razboritim istraživačkim projektima.