Intelligence as Large-Scale Computational Learning Phenomenon. (A Short Version)

Wiedermann, Jiří
1999

INSTITUTE OF COMPUTER SCIENCE

ACADEMY OF SCIENCES OF THE CZECH REPUBLIC

Intelligence as Large–Scale Computational Learning Phenomenon (A Short Version)

Jiří Wiedermann

Technical report No. 792

October 1999

Institute of Computer Science, Academy of Sciences of the Czech Republic
Pod vodárenskou věží 2, 182 07 Prague 8, Czech Republic
phone: +4202 6605 3520 fax: +4202 8585789
e-mail: wieder@uivt.cas.cz

INSTITUTE OF COMPUTER SCIENCE

ACADEMY OF SCIENCES OF THE CZECH REPUBLIC

Intelligence as Large–Scale Computational Learning Phenomenon (A Short Version)

Jiří Wiedermann[1]

Technical report No. 792
October 1999

Abstract

A coherent, machine independent computational theory of mind is proposed. The mind is presented as a specific learning algorithm whose functionality is inspired by the cognitive abilities of real brains. The respective interactive self–organizing learning mechanism, formally represented by the so–called *cogitoid,* has been recently introduced by the author.

It is shown that under a suitable training a cogitoid is able to acquire, besides abilities to handle the basic cognitive tasks, also a behavior that in its manifestation resembles that of the mind. The respective development will subsequently feature the emergence of more complex patterns of behaviour, of the self concept, and of elements of abstract thinking along with the rudimentary consciousness. Eventually, language understanding, acquisition and generation will become possible, accompanied by fully developed consciousness enabling further development of abstract thinking.

A thought experiment will show that in principle a properly trained cogitoid could pass the Turing test.

At the conclusion of the paper the implications of the previous results to various issues addressed in the philosophy of mind, such as the problem of phenomenology, computational limits to the mind's cognitive power, its efficient implementation, or its modularity are discussed. In this light also some aspects of Searle's Chinese room experiment are addressed. Finally, the problem of the future of the Turing's tests is tackled. A possible general algorithmic approach to intelligence testing is outlined without restricting it merely to human features.

[1]This research was supported by GA ČR Grant No. 201/98/0717

Keywords

cognitive computing, interactive computational learning, self–organization, cogitoid, mind evolution and functioning, consciousness, interactive Turing machine, Turing test

1 Introduction

Cognition is computation — this is a slogan whose validity seems to be increasingly more difficult to deny (cf. [4]). Nevertheless, when it comes to specifying the respective computation in greater detail, we have trouble: although apparently emerging, except vaguely described models (cf. [9]), no sufficiently complete and formal computational model of mind seems yet to exist.

A natural place where to look for computational models of whatever kind is computer science or, more precisely, its two related branches, viz. computational complexity and computational learning theory. Unfortunately, also for these fields computational brain and mind modeling is still a relatively new phenomenon. Except of a small number of initiatory works, such as [7], [12], or [21] there is not much directly related work that can be reported (cf. [25] for a recent overview of this topic). The common denominator of all the respective works is the thesis that cognition is intimately related to interactive (computational) learning. For a corresponding kind of computing whose computational mechanism is based on our ideas about brain working and whose goal is to model cognitive abilities of living organisms, the notion *'cognitive computation'* has been coined by L.G. Valiant since 1995 [22].

Building on this paradigm and on the current knowledge from AI, psychology, neurosciences, and cognitive sciences, J. Wiedermann has recently designed an algorithmic model of mind that offers quite a plausible explanatory algorithmic framework for mind evolution and its functioning [24], [26].

The purpose of the present paper is to present the first sketch of the algorithmic theory of mind which is supported by the above mentioned model. In the heart of this theory there is the statement that mind is a specific interactive learning algorithm whose self–organizing knowledge base is formed in the course of a potentially endless interaction of the mind's 'body' with the environment. This algorithm is represented by so–called cogitoids. Consciousness is an automatically emerging property of any sufficiently developed mind. It is a self–checking mechanism of mind that makes it 'fault tolerant', i.e. tolerant against its own erroneous behaviour. Consciousness thus monitors body's activities to see if actions are going as envisaged and manages the situations that cannot be handled mechanically by the previously acquired algorithmic habits. This is done by finding new, creative ways of behavior what is perceived as free will manifestation. According to their creative potential there is a continuum of minds. The model also offers a framework for discussing various philosophical issues related to the problem of computational mind.

In Section 2 we will present in a nutshell the computational model which underlies the approach at hand: the cogitoid. We will first informally explain its basic design features and show how the model responds to the basic cognitive tasks, such as simple, Pavlovian, or operant conditioning, similarity based reaction, etc. Then we will describe the self–organizing properties of its memory structures in the course of the cogitoid's meaningful interaction with its environment. At the same time, this self–structuring can be seen as the development of higher cognitive functions, such as the emergence of the abstract concepts, development of attentional and classification mechanisms, and habit acquisition. All this gives rise to a kind of an event–driven

behaviour. Under sufficiently rich external stimuli a sufficiently developed cogitoid will subsequently feature the development of language understanding and production and along with it the birth of the 'self' concept and that of the consciousness. The related notion of conscious experience is explained as well.

In Section 3 a thought experiment leading to a cogitoid's passing the Turing test will be discussed.

In Section 4 we will discuss the possible merits of the interactive learning approach to cognition to the philosophy of mind.

Finally, in Section 5 we address the issues related to the future of the Turing test as seen from the perspective of the results presented in this paper.

The idea leading to the presented computational theory of mind is based on the pioneering work by L. Goldschlager [12] and is also close to Dennett's multiple draft model [9]; the theory matches most of Dennett's ideas presented in his later book [10]and also those of M. Minsky in his 'society of mind'[16]. Of course, the model of cogitoids that incorporates recent paradigms from computational complexity theory and AI inevitably bears similarity with many other concepts in related branches of science, which are too numerous to be mentioned here (cf. [1] for recent treatments of related issues). The chapters on cognitive engines and language acquisition in Casti's book [3] present a short, readable and relatively up–to–date informal introduction into the problems treated in this paper.

2 The Cogitoid

2.1 Informal definition

When it comes to computational mind modeling we will focus our attention on programs (or equivalently, on formal abstract devices or algorithms) that are able to interactively learn, i.e. to acquire new knowledge, either all by themselves, or in a supervised manner, or in a combined way, in any situation. Moreover, this knowledge must be organized in the form of data in a way that enables the program to be 'well up' in the semantics of the respective data. By the semantics of the data we understand the relation of the data to the real events or facts they represent. Here, the events can be some external facts as perceived by system sensors, or the machine's own actions, or some internal events occurring within the program or machine. The relations can be e.g. those of similarity, contiguity in time or space, cause and effect, abstractions, etc. A general approach clearly asks for considering 'multimedia' inputs from different sensory devices. Only then can we expect that the program (or at least an intelligent agent simulating it) will understand its action in terms of the semantics anchored in real events and actions.[2]

The respective model that is conformed with the previous requirements is called *a cogitoid* and in greater detail has been presented elsewhere in [24] or [26]. Here we will only give its concise description in order to support the proposed computational

[2]Note that in our setting the predicate 'understand' has three arguments: x understands y in terms of z.

theory of mind.

The cogitoid is seen as a central part of a device that interacts with its environment. Formally, this device is represented by a kind of a transducer that transforms an infinite stream of inputs into an infinite stream of outputs. The cogitoid is equipped by potentially infinite memory in which it stores the knowledge gained in the course of its past interaction. Thus when computing the answer to the current input the cogitoid's current memory contents depending on the inputs processed thus far is taken into account

The information from and about the environment is mediated to the cogitoid via its sensors. Their task is to preprocess this information and to pass the result in the form of so-called *concepts* further to the cogitoid. The concepts are represented by the elements of some (very large) finite alphabet Γ but, on some occasions, we will find it advantageous to see them as n-tuples over the alphabet Σ (i.e. we assume that $\Gamma = \Sigma^n$). In such a representation, each element in a tuple corresponds to the presence or to a value of some feature in the description of the corresponding concept.

Based on its current input, and on the current cogitoid's memory contents, the cogitoid updates its memory and outputs some concepts. These concepts are sent to effectors for some post-processing which finally results in some action, or behavior of the cogitoid at hand.

Thus concepts are the basic entities that represent both the input (so-called *stimuli* or *perceptions*) and the output (so-called *actions*, or *behavior*) of a cogitoid. They can be seen as formal digital representations of events, of their both static and dynamic aspects, as mediated by peripherals, or of actions, behaviors, as realized by (robotic) effectors. The respective transformation from events, stimuli, perceptions into the corresponding concepts is thus the matter of peripherals. In most cases, this transformation cannot be described formally as discrete mapping since at the input to sensors we do not have a mathematical description of the perceived entity (the input can be visual, aural, or olfactory perception, or a signal from other internal body organs, of a continuous electrical, chemical, or any other nature) at our disposal. The task of the respective preprocessing is to digitalize the perceived input. The situation is similar at the output side where we do not know the formal description of (mostly motoric) actions. The cognitive system as a whole is not a pure digital system; it is rather a *hybrid system* (cf. [13] for a similar approach) that at the input and output side can make use of other than digital representations. In this way, the concepts are grounded (this is the term used by [14]) in the 'body' (sensors and effectors) controlled by the cogitoid. Concepts that are images of some external events are called *basic concepts*.

For the proper functionality of a cogitoid it is important to note that due to their very nature there hold certain relations among concepts which also hold among events or facts that are represented by concepts. These are the relation of abstraction and its inverse relation of concretization. Both relations induce a partial order among concepts. According to this order, the smaller concepts represent abstractions of the larger ones, while the larger concepts represent the concretization of the smaller ones. Mathematically, the concepts with the above mentioned operations are modeled by a finite lattice of concepts over Γ. It has the property that to any pair of concepts, a and b, there exists their greatest common *abstraction* and their smallest common

concretization in the lattice. This abstraction is given by the meet, or intersection, $a \wedge b$, while the concretization by the join, or union, $a \vee b$ of the two concepts at hand. There is one more relation defined over any pair of concepts — the relation of concept similarity. Two concepts a and b are *similar*, $a \approx b$, iff they have a nonempty intersection $a \wedge b$. The latter relation can be easily determined by making use of the tuple representation of concepts: concepts whose tuple representation coincides at some positions are similar. According to the degrees of the 'overlap' of similar concepts we can also introduce the degrees of similarity. If the previous tuples are seen as Boolean vectors and the operations of meet and join are replaced by the operations of set union and intersection then the respective lattice is a standard Boolean algebra.

In order to make a cogitoid from such a lattice, additional computational machinery realizing a certain *universal learning algorithm* is necessary.

First of all, the concepts within a cogitoid can be in two possible states: in a *passive* and in an *active state*. A currently active set of concepts, called *cogitoid's configuration*, corresponds to the single 'mental state' of a cogitoid. Initially, all concepts are in a passive state. Concepts can get activated directly from the input by so-called *external stimuli*, and from the currently active concepts, by *internal stimuli*, via so-called *associations*. Associations keep developing among the concepts automatically, as a result of interaction with the environment or as a consequence of similarity of concepts. An association between two concepts, a and b, is represented by an ordered pair of concepts $\{a, b\}$ and is denoted as $a \rightarrow b$. Thanks to its symmetry, the similarity relation, $a \approx b$, is represented by the pair of associations, $a \rightarrow b$ and $b \rightarrow a$. Each association carries a *weight* (a non–negative rational number) that is initially set to zero and is strengthened by a small amount whenever both concepts defining the association get activated. There can be *excitatory*, and *inhibitory* associations. The kind of an association that gets strengthened is determined by the *quality* of the first concept in the respective pair. The quality of active concepts is inferred from a distinguished subset of concepts called *affects*, or *operant concepts*. Affects occur in two forms: positive affects that are always of positive quality and model the *pleasurable feelings*, and negative affects that are always of negative quality and model *painful feelings*. Any other concept will inherit the quality of an active affect that is subsumed by the concept at hand. In this way, some concepts can obtain both qualities. Concepts that do not obtain a quality by these rules obtain a positive quality. The selection of concepts to be activated and selection of associations to be strengthened is described by the following computational cycle of a cogitoid.[3] Each cycle consists of four steps that are iterated all over again.

- **Input/Output**: At the beginning of each cycle, there is a set of active concepts that have been activated as the result of the previous cycle. This set forms the cogitoid's output. Simultaneously, a new set of basic concepts is activated by external stimuli. Automatically, all abstractions and concretizations of active concepts, and concepts similar to active concepts, are also activated. This corresponds to the formation of new concepts by their *simultaneous occurrence*, or

[3] In the sequel, due to the restricted space the description of a cogitoid's computational activities is somewhat simplified as opposed to the original work [24], [26] . Especially, the details concerning the initialization of concepts, the various degrees of concepts and associations strengthening and those of their gradual forgetting, are omitted.

by their *similarity*. Call the set of all currently active concepts — i.e. those activated at the end of the previous step, and those activated from the input, as $\mathcal{O}$ (stands for "old" concepts).

- **Transition to a new configuration**: Next, the excitations of all currently passive concepts, from the set $\mathcal{O}$, are computed. Here, the excitation of a passive concept is given by the sum of weights of all excitatory associations, diminished by the sum of weights of all inhibitory associations, leading from the currently active concepts to the passive concept at hand. From among all excited passive concepts the cogitoid's *selection mechanism* selects the most excited one for activation. Simultaneously with this concept, its abstractions are also activated. Call the resulting set of newly activated concepts as $\mathcal{N}$.

- **Updating the knowledge**: Now, the quality of each concept in $\mathcal{O}$ is computed and the associations between any concept from $\mathcal{O}$ and any from $\mathcal{N}$ are strengthened by a small amount. Doing so, when the quality of the first concept was of positive (negative) quality, then the excitatory (inhibitory) association gets strengthening. Otherwise, both associations are strengthened. This models the association's emergence by *cause and effect* and of *contiguity in time*. Simultaneously, all associations between the active concepts and the respective similar concepts, and vice versa, are strengthened.

- **Deactivation**: Finally, the concepts from $\mathcal{O}$ are deactivated — set to a passive state, while the concepts from $\mathcal{N}$ remain active.

Note that the sequence of configurations after each computational cycle corresponds to the 'train of thoughts' in our cogitoid. Also note that there is but a finite number of configurations, but an infinite number (countably many) of different weight arrangements. This is why a cogitoid, finding itself in the same 'mental' configuration as it did some time ago, can react to it differently — i.e. can enter a different successor configuration as it did in the past.

2.2 Basic results

The previous notion of a cogitoid can be precisely, mathematically formalized with the help of sets and mappings [24]. Within this framework rigorous mathematical theorems describing the behavior of cogitoids under a specific circumstance can be proven. This circumstance corresponds to the simplest cognitive tasks as described in the basic textbooks in psychology. First of all, any cogitoid is able to realize so-called *simple conditioning*. Confronting a cogitoid with two subsequent inputs, a followed by b, will give rise to an association $a \to b$. Providing that there are no other concepts excited from a stronger than b is, after any future activation of a, the cogitoid's selection mechanism will activate b in the next step. This will also happen in the case when a concept $a' \approx a$ is activated. Namely, by the virtue of similarity, a' will activate a which in turn will activate b. This is the basic mechanism that enables a cogitoid to behave similarly under similar situations. Learning of longer sequences of concepts is possible

as well. Pavlovian conditioning is also within the reach of cogitoids. Such a behavior is modeled first by learning the cogitoid, by a simple conditioning, an association $a \rightarrow b$. Then, the cogitoid is confronted simultaneously with a and c, followed by b. This gives rise to an association $a \vee c \rightarrow b$. Later on, when merely c is activated, and 'nothing else' represented by a special concept e, is activated in the next step, b will be invoked in the third step. The semantics of concept e is 'systems are running OK, there is nothing special to be reported'. Concept b is invoked because $a \vee c$ will invoke a by similarity and a will invoke b via the previously acquired association. After repeating such a 'cheating' (i.e. activating c instead of a) a few times, we will observe that the cogitoid ceases invoking b (this is called an *extinction*). The reason for that is that, in the meantime, the cogitoid has acquired a stronger association, $c \rightarrow e$. Hence, the selection mechanism chooses to activate e instead of b.

Note that no use of affects was needed in any of the previous cognitive tasks. This is not the case with *operant conditioning*, which deals with the acquirement of behavior, $a \rightarrow b$, whose appropriateness (or inappropriateness) is confirmed by the cogitoid's subsequent reward (or punishment). The reward (punishment) is modeled by activation of a positive affect, r (of a negative affect, p). For this mechanism to work it is necessary that the whole 'educational process' is systematically carried under the same circumstance that is called *operant context* and is modeled by simultaneous activity of the respective operant concept, denoted as k. In the case of positive reinforcement this operant concept is a positive affect, while in the opposite case, a negative affect. By the virtue of the cogitoid's computational mechanisms this context gets tied to the initially activated concept a. When activated, this concept will inherit the quality of k. Therefore, depending on this quality, there will be an excitatory, or inhibitory association established, or strengthened, between a and b. Thanks to this association, the activation of b will be enabled, or suppressed, as necessary.

Delayed reinforcement can also be handled by cogitoids.

It is important to note that all the afore-mentioned tasks can be acquired by a cogitoid mostly in an unsupervised 'training process' whose scenario is essentially determined by the environment. This scenario can freely mix the tasks, interrupt them for some time. Moreover, the training goes on also under similar circumstances. Last but not least, the same or different cognitive tasks can be 'taught' over various sets of concepts, so to speak in a concurrent way. The resulting entity is called by Dennett *Skinnerian creature* [10] since its behaviour is governed by Skinnerian (i.e., operant) conditioning.

In the sequel we will see that more complex patterns of behaviour can emerge. Namely, due to the lattice properties the concepts that have a nonempty intersection 'interact' with each other during such a process. This will give rise to a more interesting behavior that we will deal with in the sequel.

2.3 Self-organizing of the cogitoid's memory

What follows is of a somewhat speculative nature. This is because we will be interested in a complex interaction among concepts. So far, such a case seems to be beyond the limit of what is amenable to a rigorous, formal treatment. Even though no theorems

in a mathematical style can be formulated for such an interaction, we will see that the framework of cogitoids still offers a room for a plausible explanation of the underlying happenings.

When a sufficiently large cogitoid becomes a subject of a long-term, 'purposeful' interaction, in a number of increasingly complex situations, in various contexts, a certain process of self-organization of concepts will automatically start and keep running in its memory. Namely, concepts sharing the same abstraction start to 'cluster' around this abstraction. Thus a cluster with center a is a set C of all concepts that have ever been activated and $a = \wedge_{b \in C} b$. Any concept $b \in C$ is called an *episodic memory*. Each time an episodic memory, or a similar concept, is activated by the virtue of the cogitoid's computational rules (rule 1), the center of the respective cluster gets activated. Hence the centers of clusters are activated more often than the individual episodic memories. Therefore, associations leading from, and to, cluster centers are stronger than any other associations. Note also that the centers of clusters, being abstractions of episodic memories, correspond to a more general circumstance than any individual episodic memory. Therefore the clusters represent 'invariants' which hold for all participating episodic memories. Since each activation of any episodic memory activates the respective center, the whole system works as a kind of *attentional mechanism* that 'concentrates' only to features that are common to all members of a cluster. The system works at the same time as a classification system, since no matter what episodic memory is activated, the computational mechanism of cogitoids will activate the 'most similar' abstraction that subsumes the given episodic memory.

Note that clusters and their functioning resemble in many aspects the mechanism of so–called K–lines as introduced by Minsky [16].

The clusters present the basic tool for a cogitoid's understanding of semantics of individual concepts. The understanding is grounded in the basic concepts that represent directly the respective external event or actions and correspond to episodic memories.

In any cogitoid fundamental clusters and specific chains of associations evolve around three fundamental semantic categories. These categories correspond to specific operant contexts in which the interaction takes place, to objects involved in the interaction at hand, and to the way these objects are dealt with.

Contextual clusters evolve by a superimposition of episodic memories that are all pertinent to frequently occurring similar operant contexts, such as 'in the forest', 'in the street', 'Christmas', 'winter', etc. Their centers are formed by abstract concepts corresponding to objects that usually participate in these contexts. In the previous examples, it could be concepts corresponding to 'trees', 'paths', 'animals', or 'cars', 'houses', 'myself', etc. As explained in the previous part, when a particular context is activated in a cogitoid, the respective centers of contextual clusters get excited. Thus this mechanism presents a kind of an *attentional mechanism* — the cogitoid is 'reminded' of (i.e. excites concepts corresponding to) objects that used to play an important role on specific occasions. At the same time contextual clusters resemble the idea of frames as used in AI (cf.[16]).

Object clusters evolve around specific objects. The respective object presents the center of the respective cluster, while the members of the cluster provide the specific contexts, in which the object has frequently found its use in the past. A specific

object cluster will evolve e.g. around the concept 'key'. It can be used for unlocking or locking a door, a safe, a car, etc. When an object is activated, all the respective contexts in which the object at hand occurred frequently in the past will be excited. It is like offering all the possible occasions in which the object has been manipulated in the past. Thus, this mechanism presents some kind of *role assignment mechanism* for objects. To select a concrete role, additional excitation from other concepts (e.g. from consciousness — see in the sequel) is needed.

The previous two types of clusters are complemented by *functional clusters*. These are formed around frequently performed activities represented by previously mentioned specific contexts that are members of object clusters. A common abstraction of each of these activities presents the center of the respective cluster. Thus there can be e.g. a functional cluster for unlocking: its members are episodic memories for unlocking a door, a safe, etc. The respective cluster members then contain the starting operant contexts of a chain of 'algorithmic description' of the respective activities, inclusively the description of an elementary action which moves the activity towards the next step in its realization. In this sense the respective mechanism partially plays a role of the so–called *scripts* that have been known within AI for a while (cf.[16]). Since potential activities are centered around any functional clusters the cluster members also correspond to *intentions*.

Note that while the first two types of clusters — contextual and object clusters — present a kind of static descriptions free of any action, functional clusters already involve some elementary actions. To push forward the actions of a cogitoid, a specific type of its memory organization evolves along with the previously mentioned clusters. This executive part of the cogitoid's memory is given by algorithmic descriptions.

Algorithmic descriptions or *habits* are sequences of clusters are chained by associations among their centers. Each member in such sequences presents a further atomic stage in the process of realizing the algorithm at hand. By realizing one step in such a chain, the cogitoid finds itself in a new context. This new context can either activate the next step in the algorithmic chain at hand, or can trigger another activity.

Initialization of the respective chains starts at the level of corresponding basic concepts. Namely, from the computational rules described in Step 3 it follows that whenever two concepts a and b are activated in a cogitoid in two subsequent steps, an association $a \to b$ will emerge or will be strengthened. However, since both a and b are activated, all their abstractions get activated as well, by virtue of the cogitoid's computational law. Thus associations among all abstractions of a and all abstractions of b will also emerge, or will also be strengthened. This concerns especially the associations among the centers of corresponding clusters to which a and b belong. If associations among different pairs of members of different clusters are strengthened, the association among the respective centers is strengthened on each such occasion as well. It follows that the respective centers are associated stronger than any individual pair of their respective members.

Thus habits are present very strongly since they are continuously reinforced by their repeated execution under similar circumstances. Included is also an aspect of self-stimulation since the cogitoid's computational mechanism is designed so as to actively seek for opportunities to make use of habits that are appropriate for the given

occasion. On these occasions habits are continuously shaped and therefore are becoming increasingly general.

We can conclude that the behaviour of a cogitoid is driven both by the chains of acquired associations as well as by the current context in which a cogitoid finds itself. The current context activates similar, more abstract concepts that 'trigger' the respective behaviour as dictated by the chain of the respective associations. Only occasionally, at 'crossings' of some habits, the cogitoid might enter a situation where an additional input, i.e. an additional excitation from other concepts (accompanied in the typical case by the activation of consciousness — see in the sequel), is needed to direct the cogitoid's further steps. Nevertheless, under similar circumstances a cogitoid with a sufficiently evolved clusters and chains of associations will behave similarly as in the past. Even under a novel circumstance, chains of abstraction at higher levels will be found that 'match' the current circumstance and will drive the cogitoid's behaviour. Thus in practice a cogitoid can never find itself in a position when it does not 'know' what to do.

Note that in standard cases the cogitoid's behaviour will unfold effortlessly, without the necessity of making use of some inference rules.

2.4 Shaping the cogitoid's mind

The process of mind evolution can be traced in all cogitoids whose sensors are sufficiently powerful to mediate for the cogitoid a number of different basic concepts and whose effectors can realize a variety of different behaviors. Furthermore, such a cogitoid must be complex enough to accommodate a large number of concepts and of their abstractions and must be equipped by sensors and effectors that allow not only its efficient interaction with the environment, but also communication with similar species. Last but not least, such a cogitoid must undergo a slow, long-term, goal-oriented training.

The aim of this training is to learn the cogitoid to react appropriately to all stimuli. For this purpose the respective contextual, object, and functional clusters, along with the respective habits, must be established and repeatedly strengthened. Depending on the cognitive task, the training is a mix of a rehearsal, trial and error approach and supervised learning. Especially the tasks acquired by supervised learning should be taught in the order of their increased complexity. Namely, the cogitoid needs some time in order to establish and strengthen the respective habits and to link them with the current structures. It is the proper linkage of all emerging habits and clusters, their grounding in basic concepts, and their richness that take care about proper reactions to current stimuli. In this way, the basis for essential 'understanding' ('semantics') of cogitoids of 'what is going on' is prepared. So far this understanding is blind, unconscious. The abstract concepts needed for conscious understanding are present, but mechanisms for their direct activation by external or internal stimuli are not yet fully developed.

2.5 The emergence of the self concept

Among the already 'ready–made' abstract concepts there is one very special emerging distinguished concept — the *self* concept formed at the intersection of all activities governed by the mind (and its body). Hence the self concept participates in most of the concepts. In the case of animals, as a single concept it can be externally invoked by calling the animal by its name. Since the activity of this concept 'survives' from one computational cycle to the other, it participates in all operant contexts and its successive activations form a sequence to which the time arrow can be assigned. Functional clusters containing the 'self' represent activities that, according to the circumstance at hand, can or can not be activated. These potential activities may be seen as the 'self's intentions'.

The concept self alone seems to be involved in algorithmic actions that a related to the cogitoid's self–checking abilities by which it monitors its activities.

The idea about the working of the respective self–checking mechanism is roughly the following: Thanks to the rule on 'binding' simultaneously occurring active concepts the current context c_1 gets bound to the ever present self concept s and the concept $c_1 \vee s$ is activated. In the next step, an action c_2 dependent on the current context is chosen. This action is again bound to the surviving self concept and thus $c_2 \vee s$ is activated. Thanks to the similarity $c_1 \vee s \approx c_2 \vee s$, the previous context and the current one are activated simultaneously. The respective concept $s \vee c_1 \vee c_2$ (or rather its abstraction x in the appropriate cluster) will now automatically trigger the next action c_3 by following some habit. In addition to x the action c_3 also depends on the current input from sensors that 'see' the result r of the previous action. Thus c_3 can be of a 'rescue' nature, trying to remedy the result of c_2 in case that it was not as it has been 'intended', or c_3 can go along the standard line of behaviour.

In any case, c_3 must be inferred from the previous experience. As explained before, due to the abstraction potential of cogitoids some c_3 matching the current situation will be always found, but it may not be the best one for the current state of the matters. In any case, invocation of c_3 will follow automatically, effortlessly, smoothly, without any additional actions. Last but not least, if realized in the way as described before, it will be an *unconscious action.*

However, the same action can occur also in a conscious mode. This, of course, calls for explaining the emergence of consciousness.

2.6 The emergence of consciousness

Consider again the object cluster centered at the self concept. Among the episodic memories participating in the respective cluster, there will be abstract concepts corresponding to 'registering', or 'perceiving', or 'being aware', in the widest sense. In the functional cluster centered on 'registering', there will be objects that can be perceived, registered. Among these objects, there will also be the concept of the self. A prologue to the consciousness is such a configuration of the cogitoid's memory in which the 'self' excites 'observing' as its possible activity, and 'observing' excites the 'self' as a subject of observation. Since the self concept survives thanks to the cogitoid's mechanism

both concepts will become active simultaneously. In our model this corresponds to the activation of a single encompassing concept that corresponds to consciousness. Once started, the feedback between the self and the other concepts involved in consciousness will continuously strengthen the associations among the respective concepts. The habit of being conscious will emerge.

Of course, now the concept of consciousness can also take part in the previously described self–checking mechanism. In the result, all the previously described actions will be carried out 'consciously'. When compared to the previous unconscious, by the experience driven automatic self–checking, there is one important difference now. Namely, if the previously mentioned result r of action c_2 was unsatisfactory, $r \vee c_2$ can trigger an action of conscious reconsideration of the current situation. Such an action involves a kind of forecasting, introspection, and some planning.

The respective higher level mental notions can also be explained along the similar lines as before. For instance, 'forecasting' of the results of the cogitoid's specific actions is done via recalling their results from the past experience. Introspection involves the self observing (thinking about) (it)self while thinking. Both foreseeing and planning is essentially a matter of a mental simulation. It involves the self observing (it)self while 'mentally' simulating a certain train of activities. 'Mental' simulation means activation of the respective abstract concepts without letting them invoke basic concepts that control the effectors. This is achieved via inhibitory associations. The suppressing of activation of effectors must be acquired by operant conditioning.

By forecasting the consequences of possible cogitoid's actions consciousness enables to (a posteriori) compare the results of an action with the original intention, with what was envisaged or realized. Thus this implements a kind of an a posteriori self–checking mechanism.

From the above-mentioned algorithmic description of an a posteriori checking mechanism it is clear that action c_2 can be consciously checked any time after the progress report r of c_2 has been delivered. This seems to be well in line with the empirically observed phenomenon called *action–before–consciousness*. If the result was not as has been expected, then a remedy — e.g. a repetition of the original action, or a completely different action, can be consciously taken.

A more developed consciousness enables even a priori detection of situations in which the 'habit-driven' behaviour triggered by the automatic selection mechanism on the basis of the previous experience would not be beneficial. This is again done by the above-mentioned 'mental' simulation of an action (or even of a sequence of actions) to see how it could end. Note that in this way the choice of the 'best' alternative is done without actually carrying out all alternatives. In [10] the respective creatures possessing these mental abilities are called *'Popperian creatures'*. The respective process is perceived by humans as an act of *free will*, if selection is made from among some alternatives which are more or less equally good, or as an act of *planning*, if there is some prevailing alternative.

Thus consciousness is a prerequisite to the creative behaviour. In various degrees of maturity it is probably present in all but the simplest animals.

There is yet another aspect of the just described mental processes: after some time an activity requiring originally conscious processing may become unconscious. This is

the case of all activities that initially required some conscious training (such as piano playing or driving a car) but, along with the acquired proficiency, are becoming more and more automatic. In accordance with Dreyfus [11], in such a case the subject passes subsequently the stage of a novice, advanced beginner, competence, proficiency, and expert. It is the way starting with solving problems consciously, making a creative use of knowledge and experience acquired especially for the purpose at hand. At the end of this way, there is no need to solve the respective problems consciously, to make conscious decisions. The subject understands 'intuitively', what to do on what occasion, and how to do what has to be done.

The underlying cogitoid's mechanism is almost obvious: by practicing in a correct manner, the associations among the right concept activations are becoming increasingly stronger and general. The role of conscious attendance is less and less important since the new ways of behaviour are transformed into habits, 'are wired into associations', so to speak. The self is taking the role of a monitoring, self-checking mechanism. As explained before, it can work both in an unconscious and conscious mode. This algorithmic explanation is in a contrast to the main thesis of phenomenology (cf. [11]) — viz the just mentioned cognitive activities cannot rely solely on an algorithmic basis.

2.7 Conscious experience

According to Chalmers [5] conscious experience (also termed as 'phenomenal consciousness' or 'qualia') is the hardest problem related to consciousness. It is a highly subjective experience of various sensations to which there is no generally agreed-upon definition. Therefore, it is quite difficult in any model to argue that 'this is what people feel/understand as conscious experience'. In our model we can offer the following explanation of this phenomenon.

Conscious experience is equal to the activation of the respective sensual concepts accompanied by the awareness of what the sensation is about. In other words, conscious experience is a conscious understanding of what is perceived possibly attended by the activation of some affects.

In greater detail the whole mechanism works like this. Conscious understanding is caused by activating initially the concept corresponding to the sensual phenomenon. Since the consciousness is involved, the concept of the self is activated as well. The resulting concept may be appropriately called as the *self–sensual* concept. This concept will in turn activate all the respective concepts related somehow to the sensual phenomenon at hand. Essentially, these concepts are abstractions, 'invariants' of all (or the majority) of episodic memories and grounding concepts that in the past have lead to the formation of the self–sensual concept at hand. If there have been some affects involved they will also be activated to form a kind of an individual enjoyment of the sensual phenomenon. This corresponds to what is sometimes called an 'inner feel'. Of course, if there have been also some motoric activities related, they are suppressed similarly as in the case of the 'mental simulation' mentioned previously, but the inner feeling of them remains.

Altogether, these activations give rise to activation of a unique configuration of concepts (what, in our model, is equivalent to the activation of the single smallest

encompassing concept) that correspond to the conscious experience of some particular sensual phenomenon. This configuration is a highly individual one since it is based on the previous experience with the phenomenon at hand. For different species (cogitoids) living in different environments these configurations are naturally different since they are formed under different conditions. For 'non–speaking' cogitoids their conscious experience (if any) is unreportable since there is no way to express the respective information to us (people). On the other hand, for speaking cogitoids it is almost impossible to speak about the conscious experience of other species since we have no idea what it is like to be them. Therefore asking about the kind of conscious experience of the non–human mind is probably an ill–posed question.

2.8 Language understanding, acquisition and generation

When a cogitoid possesses powerful sensors and effectors that enable it to interact with its environment in an increasingly complex manner, when its memory capacity is sufficient and when subjected to the right training, a further development of mental abilities is to be expected.

Namely, the increased complexity of interaction leads to the development of an increased number of new concepts. If this is accompanied by a better mastering of, and extended sensitivity to, abstract internal stimuli, then an advanced mind evolution results.

The respective algorithmic explanation is as follows. An animal has no other than indirect means to activate certain abstract concepts. For instance, it cannot activate an abstract concept 'hunger' without being really hungry or unless seeing some food. This is because it is more or less input driven, as explained at the close of the previous paragraph, and there are no stimuli, except those mentioned, that would activate exactly and directly the abstract concept for hunger.

If there are such direct stimuli, then the cogitoid's mind would be able to treat them as any other direct stimuli. Consequently, habits, along with the corresponding attentional mechanisms, dealing only with abstract stimuli could develop in much the same way as they did in the case of concrete external stimuli.

These additional inputs that can directly activate so far unaccessible abstract concepts, are provided by the language. In the most general case a language need not be a spoken language, but for simplicity we will concentrate on this particular case. Moreover, we will consider only the case when there is already a language that a cogitoid has to learn, rather than the case when a language has to be invented (although we can treat this case as well — see the part on abstract thinking in the close of this section).

In the former case, it appears that the language to be learned must be compatible with the cogitoid's ability to generate the corresponding sounds. Generation of such sounds can be the subject of a specific training preceding that of binding the sounds to some contexts.

Namely, when a cogitoid hears a spoken language along with perceiving respective visual stimuli, by the simultaneous occurrence composed concepts consisting of words (or better: of representations of the respective sounds), and of the representation of their visual counterparts, start to emerge. By hearing the respective word the

corresponding concepts will be activated by the virtue of resemblance. The same can be achieved by pronouncing the respective word by the cogitoid itself. In the course of such a self stimulation a specific attentional mechanism will emerge, as a part of a habit that may be called 'internal speaking'. The effect of this mechanism will be that a concept can be activated without actually hearing its name. This internal activation can in turn lead to the pronunciation of the respective word, in the right operant context. This seems to be the starting point of comprehending the algorithms underlying both language acquisition, understanding, and language generation.

In cogitoids, the hearing or utterance of each word is bound to a proper *semantic* operant context that is automatically shaped in the process of language acquisition. In fact, it is the semantic operation context that provides the essential 'understanding' to cogitoids of what is spoken about. In such cases the semantic operant context can consists of complex abstract concepts reflecting the real linguistic context. What to hear and what to say in which semantic context must be acquired by a rehearsal.

Fortunately, not everything what a cogitoid can ever hear, understand, or say must be literally learned. Due to its abstraction potential, along with semantic operant contexts corresponding to the current circumstances also more abstract, *syntactic* operant concepts start emerging in the cogitoid's memory.

The syntactic operant concepts are based on the syntactic similarity of sentences. Namely, during the acquisition of a language by a cogitoid the respective abstracting mechanisms will automatically classify (i.e. 'allocate' them into the respective clusters) certain words as nouns while the other ones as verbs, adjectives, etc. Each word gets associated with the corresponding syntactic class. Moreover, by the mechanism of learning sequences the cogitoids 'discover' that in sentences the respective words usually follow the same pattern. This will give rise to syntactic operant contexts and habits that keep track of using the words in the proper order.

Both semantic and syntactic operant contexts take care of understanding and generating the language. Their proper coupling and ordering is maintained by the respective *speaking habits*, along with corresponding *semantic* and *syntactic attentional mechanisms* realized by the self–organizing cluster mechanisms. The speaking habits trigger the respective speech understanding or production frames. A kind of an *acoustic attentional mechanism* also seems to play an important role in this process.

Eventually, a picture of a complex internal network seems to emerge which is literally embodied in a cogitoid and permeates almost all of its structures. It is not 'innate', it automatically evolves in the course of the cogitoid's training and is tailored to its need. It supports both understanding and generating of a language. Its emergence and utilization by cogitoids also explains the often discussed problem of the poverty of the stimuli. This is a phenomenon that refers to the fact that, during the linguistic formative years, the child is not exposed to enough language to explain its linguistic abilities. Making use of this network one is able to generate and understand words and sentences never heard before. This idea of a network is to be compared to theories based upon assumption of some innate 'parameterized' universal grammar that fixes its parameters in the process of learning. The former idea seems to be more plausible since it is conformed with the general picture of mind development.

As the next phase of the mental development the emergence of abstract thinking

can follow.

2.9 Development of Abstract Thinking.

Abstract thinking is different from the one mostly concerning the observed world: it is a thinking about things that are non–existent and which have been invented in the process of thinking. A typical example of abstract thinking is mathematical thinking. In order that it could arise a lot more mechanisms must develop in a cogitoid than in the case of everyday thinking.

In addition to the respective abstract entities or concepts that have to be invented, defined (i.e. understand) and named in order to be able to think about them, one has to develop specific aesthetic criteria. These are defined in terms of positive or negative operant concepts whose activation motivates further abstract thinking by bringing pleasurable or uncomfortable satisfaction from it.

New rules of handling these new concepts must be invented. By their frequent 'mental' application new habits must be acquired. A specific attentional mechanism emerges corresponding to the concentration on selected issues. As a result, a corresponding 'computational' theory, with habits to think within its framework, will develop in the cogitoid's memory.

In fact, the whole process of building such a theory is not unlike the process of language invention, language understanding and language mastering: in order to think about abstract things, one has to know their meaning, to know how to deal with them, and last but not least, one has to be able to speak about them.

In this way a cogitoid can develop many different abstract internal worlds. These worlds are governed by their own rules that may or may not correspond to the observed world. Examples of such worlds span from fairy tales, fantasy, religion up to mathematical theories. As far as the latter are concerned, for instance there can be an internal world of mathematical logic, obeying the Gödel's Theorem, maintained within a cogitoid. It is clear that the limitation of this imaginary world does not concern the cogitoid's internal structure as a whole.

3 Passing the Turing test

Let us perform the following thought experiment: imagine a cogitoid being exchanged with one's brain, residing within the corresponding body. In such a case, we will assume that the cogitoid receives the same signals as the brain does. The opposite process also works: by sending appropriate signals the cogitoid can service the same peripherals as a brain does.

Then we imagine that the resulting cogitoid 'lives' in a standard human environment during a standard human life span, interacts with that environment and with other human beings.

In making this experiment, from its very beginning there is one clear advantage of human beings over our cogitoids: there seems to be a certain amount of knowledge that is somehow present in human, or in general, in animal mind without being acquired

by learning. This concerns various inherited, built–in, as it appears, instincts and reflexes, such as sucking or breathing. The corresponding activities are triggered in the appropriate situation without being ever 'trained' by the respective animal. To make the proposed thought experiment possible we will assume that cogitoids also have these innate abilities acquired by a suitable preprocessing that occurred prior the starting of this experiment.

Under such circumstances the conditions mentioned in the opening of the previous paragraph are met and therefore one can expect a gradual development of human-like mind in the respective cogitoid in the way described earlier.

Then the mind evolution can be described as a never–ending, slowly proceeding interactive learning process consisting of several phases that coincide with those described by Piaget (cf. [3]). In order to proceed to the next one, the previous one should be passed. A slight overlap in phases is possible. For space reasons we abstain from describing these phases in greater detail.

It should be clear that in this way an educated cogitoid eventually could pass the Turing's test without any principal difficulties.

4 Afterthoughts

The cogitoid, being quite a concrete algorithmic model, offers a concrete paradigm for studying, discussing, and explaining problems related to mind research. Below we shall briefly address a few of such issues.

The first note concerns the abstraction level at which cogitoids model the mind. The relation of the (human) mind to cogitoids is about the same as the relation of today's computers to Turing machines. Although any cogitoid should be able, at least in principle, to 'simulate' the mind, the respective simulation would be very cumbersome, comparable to that of real computers by the Turing machine. In any case, this possibility means that in principle both devices — the mind and the cogitoid — have the same cognitive power. This, of course, cannot be proved formally since on the left side of this equality there is a biological mechanism working on different rather than purely digital principles.

What can be done is to compare cogitoids with other standard computational models. As far as the computational power of cogitoids is concerned, it is clear that the so-called *interactive Turing machine* can simulate any cogitoid. The interactive Turing machine looks like a standard Turing machine with the following exception: in place of an extra input/output tape it has a finite number of input/output ports. At each time at most one input symbol will appear at any input port, and at most one output symbol will be sent to any of the output ports. The whole machine works as a transducer of infinite input streams into infinite output streams.

For the reverse simulation of an interactive Turing machine by a cogitoid it is necessary to equip the cogitoid with the same 'environment' and sensors and effectors as the interactive machine has. It is not difficult to show that when a cogitoid is equipped with Turing machine tapes and read/write heads it can be taught, by simple conditioning, to realize the same transition function that governs the behavior of the

Turing machine to be simulated. The question whether a cogitoid can simulate any interactive Turing machine without making use of external memories is open.

Note that by equipping a cogitoid with tapes we provide it with an 'environment' in which it can 'live' and mark places of some special–to–it interest. It reminds the Dennett's theory that humans (*'Gregorian creatures'*) 'off–load' their memory by storing symbols and hints to think and recall things in their living environment [10].

However, one should not infer from the previous remarks that a 'really working' cogitoid (mind?) can be implemented on an arbitrary universal computing device. This is because in our considerations of the mind development we have tacitly assumed that the cogitoid works in real time, i.e. that its senses, delivering its input, its computation, and its output, instructing its effectors, are all 'quick enough' to keep the pace with the changes in the environment. This is a vital condition for the proper functioning of the consciousness when seen as a self–checking system. Thus although in principle an irrigation system can perform the necessary computation, in practice it could only implement a really weird (because incredibly slow) mind. Unfortunately, almost the same holds true for today's computer systems. This is where the complexity matters are entering the game (cf. [17]).

In general, it appears that the main obstacle in realizing (simulating) the (human) mind on computers lies not in their insufficient computational power but in their insufficient computational efficiency. The just presented example of the computational mind theory shows that traditional recursive computations as realized e.g. by Turing machines are in principle sufficient. There is no need of biologically infeasible computational mechanisms, such as analog neural [19] or neuroidal [29] nets working with real numbers, or oracle Turing machines [6] possessing super–Turing computing power. The same holds true for quantum Turing machines whose computational power does not exceed that of the standard ones. Nevertheless, there is some evidence that interactive computing is in a sense indeed more powerful than the non–interactive computing [23]. Nowadays, we are probably able to computationally realize minds, but only of some simple creatures. The promising way to understand, and hopefully, in the future also to realize artificial minds of interesting size has been shown by de Bruijn in his molecular model of associative memory [8].

The efficiency barrier of cognitive computations can be overcome by their proper organization. In that respect Parberry [17] suggests an interesting idea that understanding is a method by which cognitive computations are carried out with limited resources. Cogitoids seem to be conformed with this line of reasoning.

For an overview of the computer science research agenda related to cognitive computing, see [27].

An interesting thought experiment would be to imagine a fully learned cogitoid that would be somehow 'frozen'. Such a cogitoid could perform its computations as required, but will not be able to modify its weights. It will thus remind a zombie. According to our definition of consciousness, it would not be fully conscious since it would not be able to solve creatively the 'unseen' situations. This is because to find a solution in an so far unprecedented (by a zombie) situation, the (image of this) situation must be recorded in the cogitoid's mind, in order to creatively find the proper reaction to it. Thus such a zombie cogitoid will react in a way that might not be appropriate. It

is amusing to observe that this circumstance matches well the famous Searle's Chinese room example — in this room there are no means for its operator to record some new knowledge.

There is another problem related to the Chinese room example: is it possible at all that one day the 'programers' will write a smart program, a kind of a Nuremberg funnel, which would be able to govern the behaviour of the Chinese room? Despite the complexity arguments (cf. [17]), merely in view of the previous story about the mind development in cogitoids, this does not seem to be plausible. What the programmers could do is to write some learning program, to educate it in much the same way as we did with our cogitoid, and then only 'implement' the resulting *program plus data* in the Chinese room manner. In order not to act as a zombie, Searle within the room will have a hard time in updating the respective data to reflect the new knowledge inferred in the course of conversation. Will Searle in this case understand what is going on in such a Chinese room? No, not unless the programers took care about Searle's understanding and prepared 'labels', English comments to the appropriate instructions and data manipulated by Searle. Is there ever a chance for Searle operating whatever kind of the Chinese room, to understand its behaviour? Yes, there is. The solution is simple: jointly with the Chinese room (implementing e.g. a cogitoid for that matter) Searle must undergo all the tedious process of teaching both of them in parallel, from the scratch to understand and speak Chinese. As we have seen from our cogitoid's example, this cannot be done without passing through all the phases of mind evolution. There is one additional constraint: all the sensory inputs and motoric outputs must go through Searle's body, and his brain must also participate in the experiment. Of course, in such a case, Searle himself will learn Chinese. Moreover, and this is the goal of the game, he will also understand not only what is going on in the Chinese room, but also why and how the Chinese room understands what is going on in it. By the way, he will also learn how it is like to be a Chinese.

The previous example reveals the further bottleneck of the Chinese room also observed by other authors (cf.[13]): in the mind, there is no extra 'language' module that could work independently of other parts. The mechanism of the cogitoid points to the fact that it is exactly the requirement of understanding that brings the whole cognitive mechanism into the game. That is why the Chinese room must simulate all the mind in order to be able to behave like a native Chinese speaker.

The mechanism of language understanding, acquisition and generation as presented in this paper is basically the same as that for acquiring the whole cogitoid's behaviour. There is no need for some extra, behind–the–scene 'pre–processed' structures (such as the universal grammar) as proposed by Chomsky. On the other hand, the self–structuring process occurring in the mind during the linguistic formative period of one's life can also be seen as a kind of pre–processing for (eventual) mastering of linguistic skills. But this process is of a more general nature and it also (or mainly) serves other than linguistic purposes.

In general, it is clear that modular–functional decomposition of a cogitoid is impossible: in cogitoids, everything relates to everything. The development of cogitoid's mind can also be seen as a process obeying post–modular principles of cognitive robotics [20]. In cogitoids, all these principles (called imagination, shared grounding, and incremental

adaptation in [20]) are realized by a single uniform mechanism — universal interactive learning algorithm. The 'architecture' of cogitoids thus achieves a real extreme of non–modularity — all their functionality is provided by a single 'homogeneous' component. This seems to be in accord with the 'architecture' of real brains. The presented theory of mind could thus find its use in cognitive robotics. It is of interest to compare the guiding principles of this theory with the currently evolving 'new' methodology of cognitive robotics (cf. [2], [20]). It comes as almost no surprise that both methodologies go hand in hand. Nonetheless, the presented theory goes a bit further: it offers a 'free' development of the consciousness at a level that is proportional to the overall functionality of the robot at hand and to the sophistication of its training. Therefore, what Brooks et al. call 'alternative essences of intelligence' [2] are in reality the true essences of intelligence.

The quality and quantity of the before-mentioned self–structuring process of mind leads to a wide spectrum of various minds. It starts at the level of unicellular organisms and spans over the whole animal realm, culminating with the human race. In a sense, more powerful and more efficient minds than human ones are thinkable. They will work on the same principles as the human mind, but will possess more different sensors, more communication channels, more peripherals, could have direct access to encyclopedic knowledge, could live (also) in the Internet, etc.

The last remark concerns phenomenology which claims that knowing, understanding, perceiving and the like involve more than just following the rules (cf. [3]). The example of a cogitoid points to the fact that the truth of this claim depends on the nature of these rules. If one has an old fashionable idea about once–for–all–times fixed rules that are 'ready' to handle any situation (such as the rules driving the Searle's Chinese room [18]), then the claim is correct (see the previous discussion concerning the Chinese room). However, if one speaks about the general learning rules which produce flexible, on the situation dependent dynamically changing 'second order' rules (such as the ones formed within a cogitoid), the matters look quite differently. Since the phenomenologists do not seem to explicitly exclude learning algorithms, they are on a wrong track.

The proposed computational theory of mind suggests clearly that questions asking about intelligence, and in particular, about consciousness of animals or artificial machines have no simple yes or no answers. It is rather a matter of the degree to which such entities posses the intelligence. It is to a greater extent the matter of the functionality of the brain and of its training than that of its 'architecture' or size. The above mentioned theory illustrates that rudimentary consciousness development is to be expected at quite early stages of mental development. In particular, this could be the case with animals like octopuses, chickens, dogs and the like, as increasingly often supported by the experimental evidence of animal psychologists.

5 The future of the Turing test

It has been remarked quite often that the Turing test is a test oriented exclusively towards genuine human intelligence. In principle, it will reveal the slightest deviation

towards the lower end of intelligence, from the 'standard' human behaviour.

Consider the following modification of the Turing test, aiming at the discovery of 'higher–level' learning abilities of the tested subject.

Imagine first that the person to be interrogated in the test is not a native English speaker but a foreigner speaking poor English. Now, during the conversation, such a person will ask questions such as 'would you please repeat it once more, in other words?', or ' what is the exact meaning of this word?', and the like. Depending on the language skills, the performance of a foreigner in such a test can be quite inferior. Such a subject could be easily mistakenly recognized as a computer.

Imagine now that the same person is put to the test whose goal is to discover whether he or she is able of abstract mathematical thinking. To this end the interrogator will first teach (like university students) the subject a mathematical theory unknown to it — Lobachevsky geometry, say. This might take quite a long time, but assume that both sides are patient and willing to cooperate. After some time, it will be clear to the interrogator whether the subject taught can, or cannot work within the mathematical world of Lobachevsky geometry. This conclusion can be reached despite the fact that the interrogated subject is not good at English.

The moral of this story is that a test can be constructed so as to discover the principal abilities to learn and to identify the upper bound of the subject's intelligence.

Thus any more general test aiming at discovery of intelligence in general terms should be focused towards discovering the level of intelligence on a scale to be defined. In other words such a test should discover and measure the ability of learning. Based on the algorithmic approach to the mind, tests can be designed which check the learning and creativity abilities of the subject tested. The lower end of the intelligence scale will be occupied by entities that can be subjects of a simple, Pavlovian or operant conditioning (Skinnerian creatures). It is fairly clear how to test the respective abilities. For more involved abilities (such as those of Popperian or Gregorian creatures [10]), new tests must be developed primarily by animal psychologists, taking into account the respective algorithmic mechanism and a possible algorithmic explanation. Examples of such tests include tests for the 'self' concept, conscious and unconscious self–checking abilities tests, free will, language skills, etc. These tests will not only help to reveal the mental abilities of the subject tested, but can also help in tuning the respective algorithmic model. It is beyond the scope of this paper to devise such tests, but some proposals along these lines can be found e.g. in [15].

The recent results reflecting to the unexpected linguistic abilities of apes sufficiently armed by communicating devices and properly educated suggest that in our quest for discovering the true nature of intelligence we may be quite often taken by surprise by the mental abilities of animals.

6 Conclusion

The paper reports upon work in progress. The first results and intellectual experience with cogitoids point to the fact that cogitoids or similar devices could provide an interesting framework for studying cognition. This is because they are based on general

principles consistent with the theory of animal or human psychology. Building on two basic pillars, viz. classical and operant conditioning, complemented by the mechanism of automatic abstract concept formation and reinforced creation of associations among concepts, cogitoids represent a specific form of universal learning machines. So far the results concerning higher brain function — such as mind development — are of a speculative character since we are not yet able to specify satisfactorily corresponding cognitive tasks. Nevertheless, even at this level of modeling the respective tools and results offer much more concrete framework for studying, discussing, and explaining such problems than it has been possible until now. This especially concerns the notoriously hard-to-understand problems, such as understanding of understanding, language acquisition and generation, consciousness evolution, etc. The emerging algorithmic theory of mind has been presented in this paper. A respective insight could help in devising experiments that can help in further tuning of the respective theory.

Of course, there are still a lot of open ends both in the respective computational models and the computational theory of mind. Despite of this, already at the present time does the available theory suggest that the cognition is a relatively robust phenomenon which can be explained in computational terms and is not dependent on the underlying computational model. A generalized Turing thesis seems to emerge. It states that *interactive Turing machines present a computational equivalent of an intuitive notion of cognition.* Thus it appears that the slogan from the very beginning of this paper should be improved to 'cognition is interactive computing'.

Bibliography

[1] Arbib, M.A. (Editor): The Handbook of Brain Theory and Neural Networks. The MIT Press, Cambridge — Massachusetts, London, England, 1995, 1118 p.

[2] Brooks, R.A., et al.: Alternative Essences of the Intelligence. Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI-98), the AAAI Press, 1998

[3] Casti, J.L.: Paradigms Lost. Avon Books, New York, 1989, 565 p.

[4] Chalmers, D.J.: A Computational Foundations for the Study of Cognition. Minds and machines, Vol. 4, No. 4, 1994

[5] Chalmers, D.J.: Facing Up the Problem of Consciousness. Journal of Consciousness Studies Vol. 2, No. 3, 1995, pp. 209–219

[6] Copeland, B.J.: Turing's O–machines, Searle, Penrose and the brain. Analysis, an electronic journal found at http://www.shef.ac.uk/uni/academic/N-Q/phil/analysis/preprints/preprint33.html

[7] de Bruijn, N.G.: Can People Think? Journal of Consciousness Studies, Vol. 3, No. 5/6, 1996, pp. 425–447

[8] de Bruijn, N.G.: A Model for Associative Memory, a Basis for Thinking and Consciousness. In: Proc. of the 26–th International Colloquium on Automata, Languages and Programming ICALP'99. LNCS Vol. 1664, Springer–Verlag, Berlin, pp. 74–89, 1999

[9] Dennett, D.C.: Consciousness Explained. Penguin Books, 1991, 511 p.

[10] Dennett, D.C.: Kinds of Minds: Towards an Understanding of Consciousness. New York, Basic Books, 1996, 184 p.

[11] Dreyfus, H. — Dreyfus, S.: Mind over Machine. Free Press, New York, 1986

[12] Goldschlager, L.G.: A Computational Theory of Higher Brain Function. Technical Report 233, April 1984, Basser Department of Computer Science, The University of Sydney, Australia, ISBN 0 909798 91 5

[13] Harnad, S.: Minds, Machines and Searle. Journal of Theoretical and Experimental Artificial Intelligence Vol.1, No.1, 1989, pp. 5–25

[14] Harnad, S.: The Symbol Grounding Problem. Physica D 42, pp. 335–346, 1990

[15] Kornejenko, E.: Mechanisms of Consciousness. An electronic paper to be found at http://www.glasnet.ru/~korn/c/consce.htm; the more elaborated Russian version is at http://.../consc.html

[16] Minsky, M.: The Society of Mind. A Touchstone Book, Simon& Schuster, New York, 1986

[17] Parberry, I.: Knowledge, Understanding, and Computational Complexity. In: Optimality in Biological and Artificial Networks (D.S. Lavine, W.R. Elsberry , Eds.), Lawrence Erlbaum Associates, Chapter 8, pp. 125–144, 1997

[18] Searle, J. R.: Minds, Brains, and Programs. Behavioral and Brain Sciences, No. 3, 1980, pp. 417–424

[19] Siegelmann, H.T., — Sonntag, E.: Computational power of neural networks. Journal of Computer and System Sciences, Vol. 50, 1995, pp. 132–150

[20] Stein, L.A.: Post–Modular Systems: Architectural Principles for Cognitive Robots. Cybernetics and Systems, Vol. 28, No. 6, September 1997

[21] Valiant, L.G.: Circuits of the Mind. Oxford University Press, New York, Oxford, 1994, 237 p., ISBN 0–19–508936–X

[22] Valiant, L.G.: Cognitive Computation (Extended Abstract). In: Proc. 38th IEEE Symp. on Fond. of Comp. Sci., IEEE Press, 1995, p. 2–3

[23] Wegner, P.: Why Interaction is More Powerful Than Algorithms. Communication of the ACM, Vol. 40, No. 5, May 1997, pp. 80–91

[24] Wiedermann, J.: The Cogitoid: A Computational Model of Mind. Technical Report No. V–685, Institute of Computer Science, Prague, September 1996, 17 p.

[25] Wiedermann, J.: Towards Computational Models of the Brain: Getting Started. Neural Networks World, Vol 7., No.1, 1997, p. 89–120

[26] Wiedermann, J.: The Cogitoid: A Computational Model of Cognitive Behaviour (Revised Version). Institute of Computer Science, Prague, Technical Report V–743, 1998, 17 pp.

[27] Wiedermann, J.: Towards Algorithmic Explanation of Mind Evolution and Functioning (Invited Talk). In: Proc. of the 23-rd International Symposium on Mathematical Foundations of Computer Science, LNCS Vol. 1450, Springer Verlag, Berlin, 1998, pp. 152–166 .

[28] Wiedermann, J.: Simulated Cognition: A Gauntlet Thrown to Computer Science. To appear in ACM Computing Surveys, 1999

[29] Wiedermann, J.: The Computational Limits to the Cognitive Power of the Neuroidal Tabula Rasa. In: Proc. Algorithmic Learning Theory ALT'99, to appear in LNCS, Springer-Verlag, Berlin, 1999