The Cogitoid: A Computational Model of Mind

Wiedermann, Jiří
1996

INSTITUTE OF COMPUTER SCIENCE

ACADEMY OF SCIENCES OF THE CZECH REPUBLIC

The Cogitoid: A Computational Model of Mind

Jiří Wiedermann

Technical report No. V–685

September 1996

Institute of Computer Science, Academy of Sciences of the Czech Republic
Pod vodárenskou věží 2, 182 07 Prague 8, Czech Republic
phone: (+422) 6605 3520 fax: (+422) 8585789
e-mail: wieder@uivt.cas.cz

INSTITUTE OF COMPUTER SCIENCE

ACADEMY OF SCIENCES OF THE CZECH REPUBLIC

The Cogitoid: A Computational Model of Mind

Jiří Wiedermann

Technical report No. V–685
September 1996

Abstract

A new finite formal structure — a so–called *cogitoid* — is proposed. Its computational behaviour supports a formation of new concepts and of associations among them. It is proved that cogitoids are able to realize, besides behaviour elicited by the presentation of specific stimulus–response patterns, also instances of *Pavlovian conditioning*, and even the so called *operant conditioning* that is determined by the positive or negative stimuli which occur after the responses.

Keywords

Computational Models of the Brain; Neurocomputing; Behaviourism

Contents

1 Introduction

The idea of thinking about the brain as being a computational device that can be mathematically formalized and explained, has attracted a lot of attention in artificial intelligence, cognitive sciences, and in computer science. Within the latter science we are nowadays witnessing rapid development especially in the related field of neurocomputing. Besides various kinds of neural nets designed for specific learning tasks, computational models of the brain, or mind, based on paradigms of neurocomputing, have also emerged (for a recent overview cf. [1] or [5]). Along these lines, the *memory surface model* by Goldschlager [4], and the *neural tabula rasa* by Valiant [6] seem to belong among the most elaborated models (for a brief overview of these, and other computational models of the brain, cf. [7]). Unfortunately, so far none of these or related models appears to be able to formally, mathematically treat more complicated cognitive tasks, for various reasons. In the case of memory surface this is because the Goldschlager's model is not precise enough in details that are necessary in any formal, or semi–formal reasoning. On the other hand, his model offers a good conceptual framework for the explanation of basic cognitive processes, such as concept and association formation, and even offers a plausible explanation of some higher brain functions. Contrary to this, Valiant's approach is very precise in details. It offers the notion of the so–called neuroid, a kind of a neuron that can be programmed to fullfil various atomic cognitive tasks out of which more complicated tasks ought to be assembled. At the same time this seems to introduce some limits to the potential of this model: to

explain, or model more complicated cognitive tasks at the level of actions of individual neuroids, moreover in a probabilistic setting, is not easy. A similar objection holds for other "bottom–up" approaches starting at neuronal level.

In this circumstance a sufficiently high–level model of the brain that would concentrate onto the global aspects of mental processes and would be sufficiently formal to eventually allow an exact mathematical reasoning, but at the same time would abstract from "implementation details" at the neuronal level, could be useful.

In the paper at hand a candidate for such a model — a so–called cogitoid, is presented. The basic entities with which the model deals are concepts and associations, modeled by sets and mappings. The model is not programmable in the standard sense, but can be trained to perform certain cognitive tasks by providing it with the right sequence of inputs. Any cogitoid has a spontaneous ability to create new correlated concepts by the principles of their simultaneous occurrence, or by resemblance, and to create associations among concepts that appear one after the other. Concepts are activated by external or internal stimuli and their activation can lead to formation, excitation or inhibition of other correlated concepts. A frequent invocation strengthens the presence of concepts and associations, whereas their long–term passivity leads to their extinction.

The paper is organized as follows. In section 2, main ideas behind cogitoid's definition are presented. Although these ideas are quite natural, their formal elaboration requires a certain non–trivial amount of attention to be paid to the necessary technical details. These are given subsequently in the formal definition of the cogitoid. Then the basic principles of concept and association formations are presented, and the definition of cogitoid's computational behaviour follows. The above definitions themselves are to be considered as one of the main achievements of the paper.

Next, in section 3 it is shown that the model as defined in the previous section is sound, in a certain pragmatic sense: it presents not only a suitable framework for the description and formalization of certain cognitive tasks, as described in various experiments with animals, but is able to also offer an algorithmic explanation of the respective phenomena, with sufficient accuracy. Its minimality is supported by the fact that the proofs of forthcoming theorems concerning the cognitive power of cogitoids make use of all of model's features.

The above mentioned properties of cogitoids are first demonstrated on simple cognitive tasks such as the acquisition of a behaviour elicited by the repeated presentation of specific stimulus–response patterns, or acquisition of sequences. Next, examples of standard textbook instances of Pavlovian conditioning are shown. This kind of behaviour refers to the fact that animals can be conditioned in such a way that they will tend to activate a concept that is apparently unrelated to the stimulus at hand. Final examples also give the plausible algorithmic explanation of the so–called operant conditioning. This is the behaviour that is acquired by exposing an animal to circumstances in which a certain stimuli–response behaviour is consistently rewarded or punished *after* the response is obtained. The case of delayed reinforcement is considered as well.

The underlying algorithmic justification points to the fact that behind the external manifestation of the conditioned behavior is quite a lot of internal "mental" activity which seems to challenge the generally accepted belief of behaviorism stating

that no reference to mental states in the scientific explanation of animal behavior is needed.

Finally, section 4 contains some open problems and conclusions.

2 The Cogitoid

2.1 Main Idea

The idea of a cogitoid is relatively a simple one. It is a parallel finite deterministic transducer that receives, as its input, an infinite sequence of Boolean vectors (representing concepts) of size n, and produces, as its output, an infinite sequence of Boolean vectors of size 2^n (representing actions, behavior). If properly trained and exposed to reasonable inputs its purpose is to register the frequency of occurrences of individual concepts, to perform some associative tasks over concepts, and to discover and report certain correlations among the registered concepts, such as their simultaneous or subsequent occurrence.

This is achieved quite straightforwardly by explicitly representing each concept with a respective subset of some global universum. To each concept two quantities are assigned: its strength, and its activity. The strength of a concept should measure the frequency of its occurrences, while the activity of a concept reflects the fact that the concept at hand is currently a part of the output. Moreover, there are also oriented links maintained between some concepts that indicate a correlation, or association, between these concepts. The strength of these associations is also considered. Concepts and associations may be either excitatory, or inhibitory. For most of the concepts except of concepts that are excitatory or inhibitory by their very definition, this depends on the history of their formation and/or on the current activity of other concepts.

At each moment, some concepts may be activated via the cogitoid's input, and some may be active since the previous time. These active concepts in turn activate other concepts via associative operations and associating excitatory or inhibitory links. The number of active concepts is kept bounded by the principle that only the most excited concepts get activated. These newly activated concepts contribute to the output of the cogitoid at this time, while the activity of previously activated concepts decays.

The above design of the cogitoid has been inspired, first, by principles known from experimental psychology, describing formation of concepts and associations, and second, to name only the most important of them, by principles of unsupervised learning, reinforced learning, and Darwinian selective learning (cf. [5]).

2.2 Formal Definition

Prior to giving a formal definition of a cogitoid we shall introduce some notions that will be helpful in a further explanation.

Let $\mathbb{U} = \{x_1, x_2, \ldots, x_n\}$ be a finite set of elements, called a *universum*. Then the set $2^{\mathbb{U}}$ is called a set of *concepts*. Any ordered pair (A, B) of concepts, i.e., $(A, B) \in 2^{\mathbb{U}} \times 2^{\mathbb{U}}$ will be called an *association*.

Definition 2.1 *An n–input* cogitoid *over* $\mathbb{U}$ *is a seven–tuple* $\mathcal{C} = (\mathbb{U}, \mathcal{A}, \alpha, \beta^-, \beta^+, \gamma, \psi)$ *where* $\alpha, \beta^+, \beta^-, \gamma$, *and* δ *are recursive functions, and*

- $\mathcal{A} \subseteq 2^{\mathbb{U}}$ *is the set of so–called* operant *concepts, or* affects*;* $\mathcal{A} = \mathcal{A}^+ \cup \mathcal{A}^-$, *with* $\mathcal{A}^+ \cap \mathcal{A}^- = \emptyset$. *The set* $\mathcal{A}^-$ *is called the set of* negatively reinforcing, *or* inhibitory concepts, *while the set* $\mathcal{A}^+$ *is the set of* positively reinforcing, *or* excitatory concepts.

- $\alpha, \beta^+, \beta^-, \psi$ *are functions called* memory update functions, *where*

  - $\alpha : 2^{\mathbb{U}} \times \mathbb{N} \to \mathbb{N}$ *is* the concept membership function *that assigns its* strength *in* $\mathcal{C}$ *at time* t *to each concept* $A \in 2^{\mathbb{U}}$.

    For any $A \subseteq \mathbb{U}$ *we shall say that* A *is a concept that is present in* $\mathcal{C}$ *at time* t *iff* $\alpha(A, t) > 0$.
  - β^+, *or* $\beta^- : (2^{\mathbb{U}} \times 2^{\mathbb{U}}) \times \mathbb{N} \to \mathbb{N}$ *are the* excitatory, *or* inhibitory *association functions that assign its* positive *or* negative strength, *in* $\mathcal{C}$ *at time* t *to each association* (A, B), *respectively.*
  - $\psi : 2^{\mathbb{U}} \times \mathbb{N} \to Q$ *is the* activity function, *with* $Q = \{0, 1\}$ *the set of two states, called* passive *and* active, *respectively. The function* ψ *assigns to each concept* A *its state at time* t.

- *the partial function* $\gamma : 2^{\mathbb{U}} \times \mathbb{N} \to \{-1, +1\}$ *is the so–called* quality function *that assigns its quality* -1 *or* $+1$ *to each active concept.*

The recursive definitions of $\alpha, \beta^+, \beta^-, \gamma$ *are given in the next definition of the computational behaviour of a cogitoid.*

Operant concepts, or affects, will play a key role in the computational behaviour of a cogitoid. Namely, they will be used for an *a posteriori* positive or negative reinforcement of some concepts that have been obtained as responses to some stimuli. Thus, they correspond to some specific "pleasurable", or "painful" feelings of animals (for details, see section 3.4).

To describe the computational behaviour of a cogitoid we shall introduce new notions related to principles whereby concepts and associations are formed.

According to a rather general consensus (cf. [4]), new concepts are created from simpler ones that are related by the *contiguity in space* and by *resemblance*, and new associations are formed between concepts that are related by the *contiguity in time*, or by *cause and effect.*

If A and B are two concepts, than the concept $C = A \cup B$ is called the union of these two concepts. It is formed in cases in which the concepts A and B often occur simultaneously — i.e., these two concepts are contiguous in space.

We say that two concepts A and B resemble one another in C when they share some common set C of elements, i.e., when $C \subseteq A \cap B$. This can be generalized to the resemblance of more than two concepts.

Let B be a concept. Then any concept $A \subset B$ is called an *abstraction*, or a *generalization*, of B. Similarly, any concept $C \supset B$ is called a *concretization* of B. Note that A resembles B in A and similarly B resembles C in B.

If two concepts A and B are observed in a sequence over the time, B occurring immediately after A, we speak about a contiguity in time, and write $A \rightarrow B$. This contiguity gives rise to the association (A, B) and these two concepts are then correlated via this association. This is also the case in which A is the cause and B is the effect. The respective type of associations is also called *successor associations*.

The main idea in defining the computational behaviour of a cogitoid is to provide it by the spontaneous ability to perform the before mentioned operations over concepts and to activate concepts according to their strength or according to the strength of associations among them.

Definition 2.2 *The cogitoid $\mathcal{C}$ acts as a transducer from $\{0,1\}^n \rightarrow \{0,1\}^{2^n}$ in the following way.*

The input *i_t to $\mathcal{C}$ at time $t \geq 0$ is a characteristic vector (which is an element of $\{0,1\}^n$) of some set $I \subseteq \mathbb{U}$: the j–th element of i_t is equal to 1 iff $x_j \in I$, for $1 \leq j \leq n$.*

The output *o_t at time t is the set of concepts that are active at this time: the j–th element of o_t is equal to 1 iff $\psi(O,t) = 1$, for any $O \in 2^{\mathbb{U}}$ and any $1 \leq j \leq 2^n$.*

Thus, at time $t_0, t_1, t_2, \ldots$, the cogitoid receives the inputs $i_0, i_1, i_2, \ldots$ and produces the outputs $o_0 = \emptyset, o_1, o_2, \ldots$.

At time $t = 0$, we define $\alpha(A,t) = 0$, $\beta^+((A,B),t) = 0$, $\beta^-((A,B),t) = 0$, $\gamma(A,t)$ is undefined, and $\psi(A,t) = 0$ for any $A, B \subseteq \mathbb{U}$.

Upon receiving its input, $\mathcal{C}$ performs a computational cycle *or a* step *that consists of six phases.*

In the following, let i_t be the characteristic vector of some set $I_t \subseteq \mathbb{U}$ that $\mathcal{C}$ receives as its input at time $t \geq 0$, and let $\mathcal{W}_t$ be the set of all concepts that are active at the end of the t–th computational step of the cogitoid $\mathcal{C}$.

The six phases of the t–th computational cycle are given below.

Phase 1: The Input/Output:

1. *the input vector i_t corresponding to I_t is read in;*
2. *the characteristic vector o_t corresponding to $\mathcal{W}_{t-1}$ is produced.*

Phase 2: Activating New Concepts:

The following two subphases are performed in parallel:

1. Activation from external stimuli:
   - *(a)* Activating the input concept: *I_t gets activated by setting $\psi(I_t, t) := 1$;*
   - *(b)* Activation by simultaneous occurrence: *all abstractions of I_t get activated: for all $I \subset I_t$ $\psi(I,t) := 1$;*
2. Activation from internal stimuli:
   - *(a)* Activation by simultaneous occurrence:
     All different sets of sets from $\mathcal{W}_{t-1}$ get activated
   - *(b)* Activation by successive occurrence or by resemblance:
     Let $A \notin \mathcal{W}_{t-1}$, $A \not\subseteq I_t$ be some inactive concept. The following concepts correlated to $W \in \mathcal{W}_{t-1}$ get excited, for all such W in parallel:

i. Excitation via successor associations: *Let there be an excitatory association (W, A) with the strength $\beta^+((W, A), t)$, and an inhibitory association $\beta^-((W, A), t)$. Let $S(A, t) = \sum_{W \in \mathcal{W}_{t-1}} [\beta^+((W, A), t) - \beta^-((W, A), t)]$ be the sum of strengths of all associations between A and some active concepts $W \in \mathcal{W}_{t-1}$.*

ii. Excitation by resemblance: *Let $\mathcal{R}$ be the set of all inactive concepts that resemble W in some $V \neq \emptyset$, let $A \in \mathcal{R}$. Let $S(A, t) = \sum_{B \subset A} \alpha(B, t)$ be the sum of strengths of all subconcepts of A.*

We say that the concept A gets excited with the strength $S(A, t)$ at time t.

Then the concept A gets activated at time t by setting $\psi(A, t) = 1$, iff A is a concept with the maximal excitation value $S(A, t)$ at time t among all concepts excited at time t.

When more concepts with the same excitation value satisfy the latter condition, all of them get activated.

As a result of this phase we obtain the set $\mathcal{W}_t \supseteq \mathcal{W}_{t-1}$ of currently active concepts.

Phase 3: Assigning Quality to Active Concepts:

1. Assigning quality to concepts activated in the previous cycle:
   For all $W \in \mathcal{W}_{t-1}$ do in parallel: $\gamma(W, t) := \gamma(W, t-1)$;
2. Assigning quality to newly activated concepts:
   For all $W \in \mathcal{W}_t - \mathcal{W}_{t-1}$ do in parallel: If W contains an inhibitory operant concept, or there is a concept $Z \in \mathcal{W}_t - \mathcal{W}_{t-1}$ that contains an inhibitory operant concept and $W \subset Z$, then W obtains a negative quality at time t: $\gamma(W, t) := -1$. Otherwise $\gamma(W, t) = +1$.

Phase 4: Updating the Strength:

1. Updating the strength of concepts:
   The strength of all newly activated concepts (i.e., of all concepts Z in $\mathcal{W}_t - \mathcal{W}_{t-1}$) is increased. Concepts activated directly from the external stimuli (in Phase 2.1) are strengthened more than the concepts activated from internal stimuli (in Phase 2.2):
   For all Z do in parallel:
   If $Z \subseteq I_t$ then $c := c_1$ otherwise $c := c_2$, for some constants $c_1 > c_2 \geq 1$; $\alpha(Z, t) := \alpha(Z, t-1) + c$:
2. Updating the strength of associations:
   The strength of association between all concepts $W \in \mathcal{W}_{t-1}$ and $Z \in \mathcal{W}_t - \mathcal{W}_{t-1}$, is increased, for all (W, Z). The strength of associations between concepts activated from external stimuli is strengthened more (by the constant c_1) than the strength of remaining associations (that are strengthened by c_2). Moreover, depending on quality of the first concept, either the excitatory, or the inhibitory association is strengthened. Thus:

For all (W, Z) *do in parallel:*
If $(W, Z) \in 2^{I_{t-1}} \times 2^{I_t}$ *then* $c := c_1$ *else* $c := c_2$*;*
(a) if $\gamma(W, t) = +1$ *then* $\beta^+((W, Z), t) := \beta^+((W, Z), t-1) + c$*;*
(b) if $\gamma(W, t) = -1$ *then* $\beta^-((W, Z), t) := \beta^-((W, Z), t-1) + c$.

Phase 5: Deactivation:

All concepts $W \in \mathcal{W}_{t-1}$ *are deactivated:* $\psi(W, t) := 0$.

Thus, the set of currently active concepts becomes $\mathcal{W}_t := \mathcal{W}_t - \mathcal{W}_{t-1}$.

Phase 6: Gradual Forgetting:

The positive strength of all concepts and associations whose strength has not been strengthened in the previous phases, is decreased by c_3 *with the constant* $c_3 \geq 1$.

Note that according to the definition there must be an input to a cogitoid at each time t. Nevertheless, we shall sometimes say that there is no input at time t, meaning that there is an "empty input", represented by a characteristic vector (consisting only of n 0's) of the empty set. Such an input will activate a concept called *"empty concept"* (denoted by $\emptyset$) in Phase 2. Note, however, that an empty concept is activated at each computational cycle by the virtue of Phase 2.1(b).

Further, observe that the activation function ψ described in Phase 2 plays the role of a *short–term memory*, while the functions α, β^+ and β^- play the role of the *long–term memory*. Besides, ψ models the so–called *short–term habituation*: the activation of concepts lasts only for one computational cycle, unless it is not re–invoked in the next cycle.

It is important to realize that at the beginning of Phase 2, in the t–th computational cycle of a cogitoid, two kinds of active concepts coexist simultaneously: the first kind are concepts activated from the input at time t, while the other kind are concepts in the set $\mathcal{W}_{t-1}$ that got activated during the previous cycle. Each of these two kinds of concepts activate separately, but in parallel, the correlated concepts according to the rules from Phase 2. Note that no interference between these two kinds of concepts occurs during Phase 2 which enables their independent parallel processing.

The principle used in Phase 2.2 that by successive occurence or by resemblance only maximally excited concepts from internal stimuli get activated represents some kind of *attentional mechanism*.

As a result of Phase 2, a set of newly activated concepts, denoted as $\mathcal{W}_t - \mathcal{W}_{t-1}$, is obtained at the end of Phase 2. Thus, this set contains the concepts activated from external stimuli at time t, and concepts activated from internal stimuli at time t. The internal stimuli at time t are represented by concepts in $\mathcal{W}_{t-1}$.

One can perhaps see the set $\mathcal{W}_t$ of currently active concepts as those concepts presenting the *conscious* part of cogitoid's memory. Similarly, the set of concepts that have been checked by attentional mechanism, but not activated in this computational step, can be then seen as a *subconscious* part of cogitoid's memory. Finally, the remainding concepts can present in this case the *unconscious* part of memory.

Note also the rules in Phase 3 of how the quality of concepts (i.e., the values of γ) is determined at time t. The respective function γ is defined only for concepts A active

at time t. The concepts activated in the previous cycle retain their quality. When A is activated at time t, then when it contains a negative operant concept $O \subseteq A$, or there is an active concept $B \supset A$ and $O \subset B$, then A gets a negative quality at time t.

In Phase 4, activated concepts and associations among them are strengthened. This corresponds to the experience that by the repeated activation of certain concepts the quality of their memorization keeps improving as well. Specifically, in the model the successor associations of form (W, Z) between $\mathcal{W}_{t-1}$ and $\mathcal{W}_t - \mathcal{W}_{t-1}$ are strengthened. When the quality of the concept W is positive, then the excitatory associations, otherwise the inhibitory ones, are strengthened. There may be both inhibitory as well as excitatory associations between any pair of concepts. Associations between concepts that have been activated from external stimuli are strengthened more than the other ones. This corresponds to the experience with human mind, and will be vital in explaining some cases of Pavlovian conditioning in section 3.3.

Moreover, the model works with three constants c_1, c_2, and c_3, respectively, where the first two are used as an increment in memorizing (in Phase 4), and the third one as the decrement in forgetting (in Phase 6), respectively, of concepts and associations. We shall assume that $c_1 > c_2 \geq c_3 \geq 1$ which seems to be in a good agreement with our experience.

The issue of forgetting in Phase 6 helps in freeing the cogitoid of infrequently used concepts and associations. After some time, if not refreshed, these concepts and associations will decay.

Finally, it should be observed that other choice of memory update functions is also possible. One can e.g. design a memory mechanism that would model the increasing reluctance of concepts to be repeatedly activated within short time intervals (a kind of *mid–term habituation*), or the willingness of cogitoids to prefer the activation of concepts that have been activated recently (the so–called *priming*). It is also possible to make the whole model continuous, i.e., allow for real values in the definitions of the memory update functions.

3 Simple Cognitive Tasks

3.1 Supporting Formation of Concepts and Associations

Since there is no formal specification of basic cognitive tasks, or higher brain functions, that are to be supported by cogitoids, we cannot give any definitive proof that our definition of a cogitoid is correct. The best we can do is to demonstrate that our model has abilities to model some phenomena that are generally considered to be related to mental processing of concepts as performed by human brains.

As is mentioned in section 2.2, the basic principles by which new concepts and associations among them are formed are contiguity in space and time, contiguity by cause and effect, and by resemblance. From Definition 2.2 it is readily seen that all these four principles are "built–in" directly into the transition rules of our model (in Phase 2.2).

3.2 Behavioristic Learning

The previous principles are enough to ensure the elicitation of any desired behaviour of a cogitoid solely by purposefully externally applied stimulus–response pattern via the activation of the respective concepts. Namely, by repeating the required stimulus–response pattern enough times, strong concepts representing stimuli and the respective responses are formed with strong associations among them. Then, any later activation of a stimulating concept will invoke the respective response. This can be done with many different patterns, and patterns can be presented to a cogitoid in a random order, and some of them also in parallel, grouped arbitrarily.

The details are given in the next theorem:

Theorem 3.1 *Let $\mathcal{X} = X_1, X_2, \ldots, X_k$ and $\mathcal{Y} = Y_1, Y_2, \ldots, Y_k$, $k > 0$, be two disjoint sets of concepts. We shall say that a pair $\{X_i, Y_i\}$, for any $i = 1, 2, \ldots, k$, is presented to a cogitoid $\mathcal{C}$ at time t if X_i is at the input of $\mathcal{C}$ at time t, Y_i is at the input at time $t+1$ and $\emptyset$ is at the input at time $t+2$. Let pairs be presented randomly to a cogitoid $\mathcal{C}$, in any order, with the possibility of presenting a few randomly chosen pairs at the same time.*

Then, later on, after each pair has been presented to $\mathcal{C}$ several times, whenever X_i is presented to $\mathcal{C}$ at some step, Y_i will be activated in the next step, for any $1 \leq i \leq k$.

Sketch of the proof: First assume that at time t the pair $\{X_i, Y_i\}$ alone is presented to $\mathcal{C}$ for the first time. Thus, at time t the concept X_i is activated from the input (in Phase 2.1(a)), and as a member of $\mathcal{W}_t$ it will remain active also at the beginning of time $t+1$ when Y_i will be activated from the input. Then, according to Phase 4.2 the association $X_i \rightarrow Y_i$ will be established by setting the value of the respective membership function to c_1. This association will be in the future strengthened at any similar occasion. Thus, when after a few repetitions of the above process the successor association $X_i \rightarrow Y_i$ will become the strongest one among all associations invoked by X_i, whenever X_i will be activated, Y_i will get in turn activated by the virtue of Phase 2.2(b)(i). The above arguments hold for any $1 \leq i \leq k$.

Note that when more randomly chosen pairs will be presented to $\mathcal{C}$ at the same time, then concepts encompassing all concepts appearing simultaneously will also emerge, thanks to Phase 2.1(b) and Phase 4.1. Nevertheless, the probability that such concepts will reappear in the near future is small and therefore they will tend to be forgotten due to Phase 6.

□

In a similar manner as above also sequences of concepts can be acquired:

Theorem 3.2 *Let $X_1, X_2, \ldots, X_k$, $k > 1$, be a sequence of concepts in which each concept occurs exactly once. Let the pairs $\{X_i, X_{i+1}\}$ be presented randomly to a cogitoid $\mathcal{C}$, in any order, with the possibility of presenting a few randomly chosen pairs at the same time, for any $k > i > 0$.*

Then, later on, after each pair has been presented to $\mathcal{C}$ several times, whenever X_i alone is presented to $\mathcal{C}$ at some step, for $1 \leq i < k$, concepts $X_{i+1}, \ldots, X_k$ will be subsequently activated in $k - i$ next steps, one concept at each step.

Sketch of the proof: Under the assumptions of the theorem, similarly as in the previous proof, associations of form $X_i \to X_{i+1}$ start to emerge in $\mathcal{C}$, for any $k > i > 0$. After these associations start to be strong enough to activate X_{i+1} whenever X_i gets activated, the cogitoid will start to behave as predicted in the theorem.

□

Nevertheless, more elaborated behaviour of cogitoids can be also enforced or observed as shown in the sequel.

3.3 Pavlovian Conditioning

Translated to our terminology, the Pavlovian conditioning is a phenomenon in which an animal can be conditioned to activate a concept as a response to an apparently unrelated stimulating concept (cf. [6], p. 217).

This behaviour and its variations can be precised, modeled, and explained in the framework of cogitoids in the following way:.

Theorem 3.3 *Let there be a cogitoid $\mathcal{C}$ in which such a strong association $S \to R$ has been established that S, a stimulus, alone, and only S, elicits R, a response. Let $\mathcal{C}$ undergo the* basic training process *that consists of repeating the following two steps a few times:*

- *at the beginning of a computational cycle, S appears simultaneously with another concept A at the input of $\mathcal{C}$. The concept A has so far no particular associations to R;*
- *at the beginning of the next computational cycle, there in "no input" to $\mathcal{C}$ — i.e., an empty concept $\emptyset$ corresponding to the empty input is activated;*

Then:

1. *during the basic training process $\mathcal{C}$ starts to activate R at every second computational step;*
2. *when after the end of the basic training process an input A alone is presented to $\mathcal{C}$ in the first step, and $\emptyset$ in the next two steps, $\mathcal{C}$ will activate R in the fourth step, for some time;*
3. (extinction): *after some time, instead of answering R in the fourth step, $\mathcal{C}$ will start to answer $\emptyset$;*
4. (inhibition): *moreover, when after the end of the basic training process $\mathcal{C}$ will undergo an additional training consisting of repeatedly performing the next two steps:*
   - *in the first step, an input chosen randomly from the set $\{\{A, S\}, \{A, B\}\}$ is presented to $\mathcal{C}$, where the concept B is a stimulus that does not elicit R by itself;*
   - *in the next step, $\emptyset$ is presented to $\mathcal{C}$,*

then, after some time, A alone, presented in the first step, will continue to activate R in the fourth step, while $\{A, B\}$ presented jointly, will activate $\emptyset$ in the fourth step.

Sketch of the proof: To see the first claim, note that the simultaneous occurrence of $\{A, S\}$ at the input in the first step will activate both A and S, according to Phase 2.1(b). Besides, the concept $A \cup S$ will be activated by the simultaneous occurrence (in Phase 2.1(a)) and strengthened (in Phase 4.1). In the next step, S will activate R since, according to our assumption, there is a strong association $S \to R$. Associations $A \to R$ and $A \cup S \to R$ will emerge, by the virtue of Phase 4.2. Moreover, since $\emptyset$ appears at the input, in this step the associations $A \to \emptyset$, $S \to \emptyset$, and $A \cup S \to \emptyset$ also start to establish themselves very strongly, as associations between input concepts, according to the rules in Phase 4.2. Nevertheless, for some time, thanks to the strong initial association $S \to R$, the response R will be the most excited concept and hence the cogitoid keeps activating R in every second step due to the rule in Phase 2.2.(b)(i).

As far as the second claim is concerned, observe that after some time, when the concept $A \cup S$ becomes strong enough, being repeatedly at the input, the appearance of A alone, at some time, will activate, in the second step, $A \cup S$ by the relation of resemblance (rule 2(b)(ii) in Phase 2). The latter concept will in turn, in the third step, activate A, S, and $A \cup S$. Finally, in the fourth step, the joint effort of all three latter concepts will activate R. The main contribution to this activation will come from the association $S \to R$. Here we assume that thanks to this initially strong association the excitation of R is for some time greater than that of $\emptyset$, in this step.

The extinction of R's activation will come into power when the lastly mentioned activation of R will be overruled by the activation of $\emptyset$, since the respective associations from A, S and $A \cup S$ to $\emptyset$ are repeatedly strengthened when empty inputs in the second and third steps are read by $\mathcal{C}$.

The inhibition can be explained by similar mechanisms as before. Namely, during the additional training, concepts $A \cup S$ and $A \cup B$ start to establish themselves, similarly as the associations of all combinations of concepts A, B, and S to $\emptyset$.

When A alone appears at the input, in the next step, A activates $A \cup S$ by similarity since S has a very strongly presence. In the third step S alone is activated by a simultaneous occurrence with A in the previous step (rules in Phase 2.2(a)). In the fourth step $\mathcal{C}$ eventually activates R, via the initial association $S \to R$. Here we have to assume that this association is stronger then the previously mentioned associations to $\emptyset$.

However, when $\{A, B\}$ appears at the input, A, B, and $A \cup B$ are activated in the first step. In the second step, not only $A \cup S$ is activated again, but A, B, and $A \cup B$ remain active also , due to Phase 2.2(a). In the third step, $\emptyset$ will by the most excited concept and therefore will be activated. All other concepts will be deactivated in Phase 5 of this step. Hence, in the fourth step R will not be activated.

□

Note that in order to explain the Pavlovian conditioning no use of negative operant concepts and of the related inhibitory associations was needed. Also observe that in the case of extinction and inhibition the cogitoid keeps activating some concepts

in its second and third step, but others than R.

3.4 Operant Conditioning

Now we show that our model is also able to realize so–called *operant behaviour*. This is a behaviour acquired, shaped, and maintained by stimuli occurring *after* the responses rather than before. Thus, the invocation of a certain response concept R is confirmed as a "good one" (by invoking the positive operant concept P) or "bad one" (the negative operant concept N) only after R has been invoked. It is the reward (P), or punishment (N) that act to enhance the likelihood of R being re–invoked under the similar circumstances as before.

The real problem here is hidden in the last statement which says that R should be re–invoked (or not re–invoked) only under similar circumstances as before. Thus, inhibition, or excitation of R does not depend on S alone: in some contexts, R should be inhibited, and in others, excited.

As an example, think of the idea of adding spice (say, a chilli paprika) to some food. While in some cases it might be a good idea (like in chilli con carne), in some others it could be complete nonsense (like in the case of an apple pie, say). Thus, if S is the concept related to food, and R to the adding of chilli paprika, in the former case R should be positively reinforced and eventually activated, while in the latter it should be definitively inhibited and eventually not activated.

The implementation of such behaviour is described in the following theorem.

Theorem 3.4 *Let $\mathcal{C}$ be any cogitoid. Let $t_1 < t_2 < \ldots$ be the distinct times, with $t_{i+1} > t_i + 3$. Let, at these times, $\mathcal{C}$ to find itself in the same so–called* operant context *$\mathcal{Z}$, where $\mathcal{Z}$ is the maximal subset of active concepts in $\mathcal{C}$ that remains invariant at the beginning of times t_i, $t_i + 1$, $t_i + 2$, and $t_i + 3$, for $i = 1, 2, \ldots$.*

Let a strong association $S \to R$ be established in $\mathcal{C}$ prior to time t_1. Let R be such a response that, whenever activated simultaneously with any $\mathcal{Z}$, it will be consistently punished by internally activating a negative operant concept N (consistently rewarded by a positive operant concept P) in the next step.

Then, whenever the stimulus S alone is presented to $\mathcal{C}$ at some time $t_i + 1$, then

- *when R is to be rewarded, R will be activated at time $t_i + 2$, for $i = 1, 2, \ldots$*
- *when R is to be punished, then there exists a constant $k > 0$ such that for $i = 1, 2, \ldots, k$, S elicits R at time $t_i + 2$, whereas for $i = k, k + 1, \ldots$, S will not elicit R at time $t_i + 2$.*

Prior to giving the proof of this theorem, a few words to explain its essence are in order.

Observe that the assumptions of the theorem are fairly general. To take a lesson from behaviour confirmed by an operant concept, no particularly intensive, continuous training is needed. It is enough when a similar circumstance, modeled by the concept S and the operant context $\mathcal{Z}$ (like adding chilli to various kind of fruit pies) will occur a few times (i.e., at times $t_1 < t_2 < \ldots$), possibly with long intervals between

the successive occurrences. When the subsequently elicited action R will be always punished, the theorem says that after only a few punishments the cogitoid ceases to invoke R at these occasions. The "similarity" of circumstances is modeled by the requirement that the context $\mathcal{Z}$ remains the same at all pertinent occasions. This is fulfilled e.g. in cases when, at all times, $\mathcal{C}$ finds itself in configurations that share a subset of sufficiently strong concepts that alone uniquely characterize the event (to return to our running example, such a set may be the concept of a fruit pie, of which an apple pie is a concretization).

Sketch of the proof: The idea of the proof is as follows: we have to achieve that when at each time $t_i + 1$ after the time t_1 whenever the concepts in $\mathcal{Z}$ are activated and S appears at the input, $\mathcal{C}$ has to "recall" what it did at a similar occasion in the past, i.e., at time t_{i-1} and shorty afterwards. Then it has to accept the decision that was either approved by P (i.e., reinforce the excitation of R), or rejected by N (inhibit the excitation of R) at time $t_{i-1} + 3$.

In the sequel we shall describe the actions of cogitoid $\mathcal{C}$ under the circumstances as described in the theorem. We shall mainly concentrate on the description of the most important events which will happen in the cogitoid that will be crucial to prove our theorem.

Consider the state of affairs in our cogitoid at time t_1.

At this time, the operant context $\mathcal{Z}$ presents the set of concepts activated at the end of the previous step.

At time t_1+1, S appears at the input and $\mathcal{Z}$ is still active, due to our assumption. Subsequently, in the next step, the concept $S \cup \mathcal{Z}$ will be created, R will be activated, and the association $\mathcal{Z} \to R$ will be created.

In turn, at time $t_1 + 3$, internal activation of an operant concept O will follow, according to our assumptions. Thanks to our assumption on $\mathcal{Z}$, this concept will still be active, and thanks to its simultaneous occurrence with O, a composed concept $\mathcal{Z} \cup O$ will appear. Also the association $\mathcal{Z} \to \emptyset$ will be formed.

Next time, at time t_2 the cogitoid $\mathcal{C}$ will be again in the context $\mathcal{Z}$. In the next step, at time $t_2 + 1$, $\mathcal{Z} \cup O$ will be activated. Now, when stimulus S also appears at the input, it will activate the response R in the next, $t_2 + 2$–nd step. Nevertheless, at this occasion the association $\mathcal{Z} \cup O \to R$ will be also established, with the same quality as O has, due to the quality inheritance rules in Phase 3. Moreover, the concept $\mathcal{Z} \cup O \cup S$ will be created for the first time.

The latter mentioned association $\mathcal{Z} \cup O \to R$ will be repeatedly strengthened at times t_i+1, for $i = 3, 4, \ldots$. Note that at time t_3+1 even the association $\mathcal{Z} \cup O \cup S \to R$ will emerge which will then also contribute to the negative or positive excitation of R. When O was a negative operant concept, from a certain time $t_k + 1$, both associations will become strong enough to jointly prevent the activation of R. Clearly, when $\mathcal{Z}$ is not active, there is no inhibitive contribution to R and therefore the activation of R can be still elicited by first activating S.

□

From the previous theorem and from its proof the following interesting fact follows. When, at time $t_1 + 2$, the response R is activated and it is a correct one, confirmed by P, the cogitoid behaves like having already acquired the right behavior.

But from the proof it is seen that this is not yet the case, since R is activated by the stimulus S alone, not yet taking P into account. In the case of an incorrect answer the cogitoid takes a proper lesson very fast and after a few exercises behaves "correctly". Note, however, that due to Phase 6, in the long run the cogitoid tends to forget what it has learned in this case, since the inhibitory connections are not strengthened any longer, because R is not invoked.

In practice this means that behaviour acknowledged by positive rewards looks like it is being acquired faster than the one acknowledged by negative rewards. Moreover, in the long run, if not rewarded, the inhibited behaviour tends to be forgotten. This seems to correspond well to our everyday experience — reinforcing good habits is easier than suppressing wrong ones.

3.5 Delayed Reinforcement

It appears that by a similar mechanism that ties a certain operant concept to some temporarily prevailing operant context one can also explain a more complicated case of the so–called *delayed reinforcing* when the reinforcing stimulus — a punishment or a reward — does not necessarily appear immediately after the step to be reinforced.

Thus, we shall deal with a problem when a punishment appears only after the cogitoid activates a previously acquired sequence of concepts in a way described in Theorem 3.2.

Theorem 3.5 *Let a sequence of concepts $X_1, X_2, \ldots, X_k$, $k > 1$, in which each concept occurs exactly once, be acquired by a cogitoid $\mathcal{C}$. Let i be fixed, with $1 \leq i < k$, let $\mathcal{Z}$ be an operant context that is active only during the activities of $X_i, \ldots, X_k$.*

Then, if each activation of X_k subsequently invoked by the initial activation of X_1 (cf. Theorem 3.2) will be punished by activating a negative operant concept N, then, after this happens a few times, activation of X_1 will activate $X_2, \ldots, X_i$, but none of X_j, for $i < j \leq k$.

Sketch of the proof: The similar mechanism as in the Theorem 3.4, viz. the repeated punishment of X_k by N in the context Z, will give rise to the negatively reinforcing association $\mathcal{Z} \cup N \cup X_{k-1} \rightarrow X_k$. Thus, under the circumstances at hand after several repetitions of the previous experience $\mathcal{C}$ ceases to activate X_k, but will still activate X_{k-1} (when $i < k-1$). Nevertheless, the latter activation will by resemblance activate the concept $\mathcal{Z} \cup N$ that has emerged previously (see the proof of theorem 3.4). This concept will act like a punishment for the activation of X_{k-1}, and hence the negatively reinforcing association $\mathcal{Z} \cup N \cup X_{k-2} \rightarrow X_{k-1}$ will eventually emerge. After some time, it will start to inhibit the activation of X_{k-1}. The similar process repeats itself for X_{k-3}, etc., until it reaches X_i. Then, clearly, in accordance with Theorem 3.2 activation of X_1 will still activate the concepts $X_2, \ldots, X_i$. Nevertheless, unlike the preceeding ones, the latter concept, X_i, will be active simultaneously with the operant context $\mathcal{Z}$ whose association with the negative operant concept N will cause that none of X_j, for $i < j \leq k$, will be activated.

□

Based on the proof of the previous theorem it appears that cogitoids are also able to realize a strategy that is similar to the so–called *backtracking* technique used in searching for the accepting solutions in decision trees. In this case, at each potential branching point cogitoid must find itself in a specific operant context. When some sequence of concept activation presents a "blind alley" (i.e., leads to a punishment) in this context, this operant context will get associated with the negative operant concept much in the same way as in the proof of theorem 3.5. This will subsequently enable the cogitoid to return to the branching point and, eventually, to take an other path. However, ve will abstain from stating the respective theorem since its proper formulation requires a lot of technical assumptions.

3.6 Behaviorism and Mental States

It is generally accepted (cf. [2] or [3]) that behaviorism is a school of psychology that is being built on a firm belief that no reference to mental states in the scientific explanation of animal behaviour is needed.

This belief seems to be questioned in the context of cogitoids. Namely, especially from proofs of theorems 3.3, 3.4, and 3.5 it is seen that in order to realize the conditioned behaviour a lot of internal, "invisible" activities in the respective cogitoids have to take place. It is a matter of taste whether we shall identify the respective configurations through which the cogitoid has to pass with mental states or not.

But if we do so then the cogitoid must be taken as a plausible model of mind. Then also the above behavioristic belief should be revised.

4 Conclusions and Open Problems

The contribution of this paper is twofold. First, a new formal finite model of mind has been introduced. It presents a deviation from similar models that are based on paradigms of neurocomputing. This is because it abstracts from brain–like implementations and instead focuses on the substance of mental processes — viz. the concept and association calculus. Secondly, the viability of this model has been demonstrated by its capability to exactly formulate the instances of Pavlovian and operant conditioning and to give their plausible algorithmic explanation. This seems to be the first occasion when a computational justification of the related mental phenomena, in a form usual in computer science, is given.

Of course, the underlying emerging theory of cogitoids has a lot of open ends. First of all, the versatility of the model must by further verified on other cognitive tasks described in experimental psychology. This will probably lead to further tuning of the model. But already now, in its present form the model seems to offer an interesting and promising alternative for studying the computational aspects of mental processes. It opens ways for its possible computer simulation. It appears that with the help of this model algorithmic explanation of higher brain functions (such as the concept of self, consciousness and unconsciousness, thinking, language acquisition and generation, etc.) could be also possible. Such possibilities are subject of the author's current research in

this field. The work by [3] or [4] can serve as a vital source of inspiration along these lines.

Acknowledgements Thanks to Hava Siegelmann and Pekka Orponen for the enlightening discussions related directly or indirectly to the subject of this paper during my visits to Technion, Haifa, University of Helsinki and University of Jyväskylä.

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

[3] Dennet, D.C.: Consciousness Explained. Penguin Books, 1991, 511 p.

[4] 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

[5] Haykin, S.: Neural Networks: A Comprehensive Foundation. Macmillan College Publishing Company, Inc., New York, 1994, 696 p.

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

[7] Wiedermann, J.: Towards Computational Models of the Brain: Getting Started. Technical Report No. V–678, Institute of Computer Science, Prague, June 1996, 33 p.