The Natural-Constructive Approach to Representation of Emotions and a Sense of Humor in an Artificial Cognitive System

Olga Chernavskaya
Laboratory of elementary particles
Lebedev Physical Institute (LPI)
Moscow, Russia
e-mail: olgadmitcher@gmail.com

Yaroslav Rozhylo
BICA Labs
Kyiv, Ukraine
e-mail: yarikas@gmail.com

***Abstract*— The Natural-Constructive Approach is proposed to describe and simulate the emotions and a sense of humor in an artificial cognitive system. The approach relates to the neuromorphic models and is based on the original concept of dynamical formal neuron. The main design feature of the cognitive architecture consists in decoupling the cognitive system into two linked subsystems: one responsible for the generation of information (with the required presence of random component usually called "noise"), the other one – for processing the well-known information. The whole system is represented by complex multi-level hierarchical composition of neural processors of two types that evolves according to certain principle of self-organization. Various levels are shown to correspond to the functional areas of the human-brain cortex. Human emotions are treated as a trigger for switching the subsystem activity that could be imitated and mathematically expressed as variation of the noise amplitude. Typical patterns of the noise-amplitude variation in the process of problem solving are presented. The sense of humor is treated as an ability of quick adaptation to unexpected information (incorrect and/or incomplete forecast, surprise) with getting positive emotions. Specific human humor response (the *laughter*) is displayed as an abrupt "spike" in the noise amplitude. Thus, it is shown that human emotional manifestations could be imitated by specific behavior of the noise amplitude.**

Keywords- noise; emotions; explanatory gap; spike; surprise.

I. INTRODUCTION

Recently, the paper concerning the interpretation of emotions and the sense of humor in an artificial cognitive system was published and presented at the conference COGNITIVE 2016 [1]. This paper represents an invited extended version.

The problem of modeling and imitation of the cognitive process is actual and very popular now, especially in the context of Artificial Intelligence (AI) creation. Among the most popular approaches, there are Active Agent paradigm (e.g., the SOAR architecture, see [2], [3]), Deep Learning paradigm [4]–[6], Brain Re-Engineering [7], [8], Robotics [9], Resonance theory [10], etc. The majority of imitation models proposed are aimed to construct the artificial cognitive systems for solving certain (even broad) set of problems *better* than human beings do. Hence, those systems have to be *efficient*, *reliable*, and *fast-acting*.

In our works [11], [12], so called Natural-Constructive Approach (NCA) has been elaborated, which is focused on modeling just the human-like cognitive systems. Therefore, the priority is given to the features inherent to the *human* cognition, such as *individuality*, *intuitive* and *logical* thinking, *emotional impact* on cognitive process, etc. This approach is based on the Dynamical Theory of Information [13]–[15], data from Neurophysiology [16]–[18], and Neuropsychology [19], and Neural Computing [20]–[22] (with the latter being used in a modified form). Note that NCA could be related to the Human-Level Artificial Intelligence (so called *HLAI* track, see, e.g., [23]) and is close to some extent to the Deep Learning paradigm [4]–[6], but possesses certain important and original peculiarities presented below.

This paper is focused on modeling the manifestation of emotions in the cognitive process. The version of the human-like cognitive architecture elaborated under NCA is presented schematically. The main constructive feature of this architecture consists in *decoupling* the cognitive system into two linked subsystems: one responsible for generation of information (with required presence of random component, i.e., "noise"), the other one — for reception and processing the well-known information**.** The activity of these subsystems is proposed to be controlled by the *emotional* mechanism. Switching the subsystem activity is associated with the noise amplitude variation, which could be related to the change in neurotransmitter composition. This paradigm is applied to simulate the human reactions under stress conditions (including "smooth" stress, i.e., surprises). A particular case of the noise-amplitude behavior, — namely, the abrupt up-and-down change ("spike"), — is treated as an analogue to human *laughter*.

The paper is organized as follows. Section II presents a brief overview of modern approaches to representation of emotions in AI. Section III describes basic components of NCA. Section IV describes the main constructive blocks of cognitive architecture designed under NCA. In Section V, we discuss the role and place of emotions in the proposed architecture and present the example of application of the proposed model to describe the effects of stress/shock. In Section VI, typical manifestations of emotions in course of solving different problems are considered; special attention is paid to representation of the sense of humor in AI. Perspectives on practical validation of the results obtained

are discussed in Section VII. In Conclusion, main results are summarized and future perspectives are discussed.

II. MODERN STATE OF THE EMOTION REPREENTATION PROBLEM

Simulation of the human-like cognitive process implies inherently the integration of rational reasoning and emotions into one cognitive system. This problem represents one of the main challenges as for *AI*, as well as for any human-level cognitive architecture (*HLAI*, [23]). The main problem here is connected with the so-called "explanatory gap" [24], i.e., the gap between the "Brain" (cortical and sub-cortical structures) and the "Mind" (consciousness). This means that there is a lot of information from the B*rain* side (neurophysiology) on the structure and functions of single neuron, and even on the neuron ensemble (e.g., [14]). On the other ("Mind") side, there is a lot of information from philosophy and psychology (including personal experience) on the consciousness manifestations (e.g., [25], [26]). However, there is a lack of ideas on how the first could provide the second.

A particular consequence of this fact is surprisingly poor and vague definitions of such concepts as *emotions*, *intuition*, *logical thinking*, *subconscious*, etc., which are presented in such respective Dictionaries as Miriam-Webster [27]. However, definitions from the Wikipedia [28] seem more meaningful, modern, and reasonable in our view.

The same "gap" concerns as well the representation of emotions. On the "Mind" side, emotions represent, according to definition "…*subjective self-appraisal of the …current/future state*" [28]. On the other side, from the "Brain" viewpoint **(**see, e.g., [7], [16]**)**, emotions are treated as a *composition of neurotransmitters* produced by certain sub-cortical structures. This value is objective and experimentally measurable. But where is the "bridge" between the neurotransmitter composition and personal feeling of satisfaction, disappointment, etc. — that is the question.

This problem actually attracts attention and evokes a lot of studies (see, e.g., [29] –[39]). However, the variety of approaches to the problem of emotion representation indicates itself that the problem is not solved yet, so that, "…*emotions still remain an elusive phenomenon*" [40].

Below, we try to collect the interpretations and main features of emotions provided by different approaches and propose our view on accounting for emotional component in the artificial cognitive system.

The approaches from the "Brain" viewpoint refer mainly by the Brain Re-Engineering paradigm (e.g., [7], [8], [29], [30], [31]). It is based on the analysis of complementary role of cerebral cortex and certain sub-cortical structures — *thalamus*, *basal ganglia*, *amygdale*, etc., — directly related to the control of the emotions in cognitive process. This way looks very close to the goal, but the consideration actually seems mostly verbal: the mathematical apparatus used seems rather poor. Moreover, the role of emotions is attributed mainly to the *reinforcement learning* process, while it is important but far not the only act of cognition. Besides, these studies focused on the motor (acting) training, leaving aside the cognitive process itself.

Another, somewhat more abstract "Brain-inspired" approach is presented by the works of Lovheim and followers (see [32], [33]). Here, the three-component model was proposed that involved three systems of monoamine neurotransmitters (namely, *serotonin*, *dopamine*, and *nor-adrenaline*), which provide cubic representation of various emotional states. This model is popular and provides good results for describing several medical problems (deceases), but seems not so well in modeling regular cognitive process.

From the "Mind" viewpoint, the majority of researches refer to the *active agent* concept ([2], [34], [35]). Here, the agents are supposed to have the ability of self-appraisal from the very beginning, and the question is: how this appraisal does influence their reasoning. There were suggested various principles of organization of the "emotional space" that affect the cognitive process. However, the main problem from our viewpoint is to understand the very mechanism that could provide the self-appraisal ability.

A similar way is to introduce several discrete emotional states that would affect (with certain weight coefficients) the model calculations for *AI*. Their number may vary — from two (positive and negative ones) up to 27 in [34]. However, clear mechanisms of emotion emergence are not revealed in any of these cases.

The other approach ([36], [37]) involves two sets of dynamical variables, emotional and rational ones, so that their (nonlinear!) interaction results in various states of the system providing certain nontrivial regimes of transition between those states. However, the neurophysiology interpretation of the emotional, as well as rational, variables under this approach remains somewhat dissatisfactory.

An interesting (but somewhat shocking) idea was put forward by Schmidhuber [38]: the ultimate goal of living activity that provides the most positive emotions is connected with the *compression* of information. Being seemingly not the most actual goal for a human being (as compared with, e.g., *survival*), it could be reformulated in terms of "image-to-symbol conversion" (see below). Then, this idea surprisingly meets our final inferences.

The last but not least, let us turn to the concept suggested by Huron [39] that emotions are evoked by *anticipations*. In spite of this hypothesis is formulated rather verbally than mathematically, it seems the most promising and could serve as a basis for mathematical modeling.

Note that common modern trend consists in associating emotions not with *particular state*, but with certain *transitions* between different states (see [35], [39]). This trend seems to be the most promising since it does not fix or limit the number of mechanisms (as well as neurotransmitters) that provide emotional manifestations, but is focused on the *variability* of the cognitive process.

This study represents an attempt to merge the "Brain" and "Mind" paradigms under NCA by revealing (or introducing) proper variables and coupling them into unified dynamical system (i.e., "emotional block", see below).

III. BASIC COMPONENTS OF NCA

The approach NCA is aimed to understand and reproduce in mathematical model the human-like cognitive features like *spontaneity*, *paradoxicality* (the ability to formulate and solve paradoxes), *individuality*, *intuitive* and *logical* reasoning, integration of *emotions* and rational reasoning. Therefore, certain paradigms typical just for the living objects are required. NCA involves one of such paradigms provided by the Dynamical Theory of Information.

Being biologically inspired, the approach belongs to so called neuromorphic models, which implies that the neuron is the basic element (in some sense, the "*active agent*") of the whole cognitive architecture. Hence, both neurophysiology and neuropsychology data should be taken into account.

Neural computer paradigm is used for computation and numerical simulations. Under NCA, somewhat modified representation of the neuron that was called the "*dynamical formal neuron*" model is employed.

Thus, NCA combines actually three areas of expertise.

A. Dynamical Theory of Information

The Dynamical Theory of Information (DTI) is relatively new theory elaborated in the post-middle of XX*th* century, almost at the same time as the well-known theory of communication of Shannon (see [41], [42]). However, Shannon's theory was focused on the process of information transmission, while DTI analyses the process of its origin and evolution. This theory, being the subfield of Synergetics (see [13], [43]), was elaborated in the works of Haken [13] and Chernavskii [14], [15]. It is based on the idea that the information is a specific kind of object that possesses simultaneously as solid (material), as well as virtual features. The information appears as a result of evolution and interaction within certain community of living subjects. Let us stress that the brain, being an ensemble of neurons, represents a specific case of such community.

The most constructive and explicit definition of Information belongs to Quastler [44]: "*The Information is the memorized choice of one version of N possible (and similar) ones"*. This definition provides immediately the possibility to reveal different types of information:

- *Objective* Information — the choice done by the Nature as a result of its evolution, i.e., physical (objective) laws reflecting real structure of the surrounding world.
- *Conventional* (*Subjective)* Information **—** the choice done by a group of subjects as a result of their interaction, communication, fight, agreement, convention, etc., that is individual for a given community.

In the first (Nature) case, the choice appears to be done according to the principle of minimum energy expenses. In the second (people) case, the particular choice should not be *the best* one, but should be done and stored. The most widely-known examples of conventional information are the following: language, alphabet, traffic signs, symbols, etc. A particular language could be neither better nor worse than other, but it reflects the *mentality* (individuality) of a given society (see, e.g., [45]).

Moreover, that definition provides the idea of *how* the information could emerge. There are two mechanisms:

- *Perception* — superimposed (externally forced) choice associated with the Supervisor learning.
- *Generation* — free (random) choice that should be done without external control (internally).

It was shown in [13]–[15], that the information generating process requires mandatory the participation of chaotic element (so called "mixing layer") that is commonly called the *noise*.

The main inference of DTI is that these two mechanisms are *dual* (or *complementary*), and hence, *two subsystems are required to perform both these functions*. In analogy with two cerebral hemispheres of human brain, let us call these subsystem Left Hemi-system (**LH)** and the Right Hemi-system (**RH**), respectively.

From the positions of DTI, the *cognition* is considered as a process *of processing the information.* Therefore, the cognitive process could be defined as "*the self-organizing process of recording (perception), memorizing (storage), coding, processing, generation and propagation of the personal conventional information"* [11]. Note that this definition does presume the *subjective* (individual) character of human thinking.

B. Neurophysiology and Neuropsychology Data

Let us stress that both, the "Brain" and the "Mind" evidences should be taken into account. "Brain" data concern the neuron structure and mechanisms of their interactions.

1) Neuron Representation: NCA refers to so called "neuromorphic" models. This implies that the basic element for any structure is the neuron. In neurophysiology (see, e.g., [46]), the neuron model presented by Hodgkin-Huxley [47], as well as its somewhat reduced version suggested by FitzHugh-Nagumo [48], [49], are considered still as the most relevant ones. Starting from the Fitz-Hugh model, we have elaborated the *dynamical formal neuron* concept (see [11]) that represents a particular case of this model. Accordingly, nonlinear differential equations were used to describe the single-neuron behavior and the neuron interactions. This enables us to trace the dynamics and reasons for symbol formation.

2) Neuron Interaction Representation: Experimental data on interaction in the neuron ensemble show:

a) Numerous experiments indicate that the perception of new information is accompanied by amplification of the connections between neurons involved in this process. This is called the "Hebbian rule" [17].

b) Modern experimental data on the neuron structure [18] show very intriguing fact: those neurons that participate in acquiring certain experience ("skill") appear to be modified as compared to free (unemployed) neurons. This inference is based on the experimentally observed

distribution for the expression of so called c-FoS gen responsible for changing the neuron structure. Thus, the proper model representing a neuron should involve the possibility of a certain mutation for engaged (trained) neurons.

3) Neuropsychology Evidence: Another challenge for any relevant model of a human-level cognitive system is the question: why there are just two cerebral hemispheres in the human brain – the right (**RH**) and the left (**LH**) ones. From psychological viewpoint, we take into account the wide-spread opinion that **RH** is associated with non-verbal, imaginary, *parallel* thinking and intuition (see, e.g., [26], [50]). Correspondingly, **LH** is associated with *sequential verbalized* thinking and the logical reasoning. However, while there is no clear explanation of intuition and logic, these statements seem ambiguous.

Another, more constructive from our viewpoint, idea had been put forward by E. Goldberg (practicing psychologist) [19]. He infered that **RH** is responsible for processing *new* information, i.e., learning, while **LH** has to process the *well-known* information. Note that this concept entirely coincides with the main inference of DTI, that any cognitive system should contain two subsystems, one for generation of new information, the other one for reception and processing the existing information.

C. Neurocomputing

A cognitive system could be presented as a composition of neural processors, i.e., the plates populated with model neurons. It should be stressed that, in contrast to common neural computing (see, e.g., [51]) based on the simple formal neural paradigm suggested by McCulloch and Pitts [52], NCA is based on the concept of *dynamical* formal neuron presented in [11].

Two types of neural computers are employed:

1) Distributed memory:

This concept refers to the Hopfield-type processor (***H***) with *cooperative* intra-plate ("horizontal") interaction [20]. Any real object is represented as a "chain" of activated (excited) neurons, which is called the "*image"* of this object. The main advantage of such type of representation is connected with the fact that the damage of few neurons of this chain does not lead to the damage of the image as a whole. The integrity of the image is secured by trained connections between the neurons involved into the image formation.

Note that real objects having similar fragments are to be written by the *overlapping* chains of neurons, which provide *associative* connections between these objects.

The model of the ***H***-type processor with dynamical formal neurons could be written in the form:

$$\frac{dH_i(t)}{dt} = \frac{1}{\tau^H}[\{H_i - \beta_i \cdot (H_i^{\,2} - 1) - H_i^{\,3}\} + \sum_{i \neq j}^{n} \Omega_{ij} \cdot H_j] \equiv \frac{1}{\tau^H}[\Im_H\{H_i, \beta_i\} + \sum_{i \neq j}^{n} \Omega_{ij} \cdot H_j] \quad (1)$$

where $H_i(t)$ is variable describing the state of i-th dynamical formal model neuron, τ_i^H— activation characteristic time, β_i — parameters that characterize the neuron excitation threshold. The functional $\Im_H\{H_i,\beta_i\}$ describes the internal dynamics of a single H-type neuron, the second term refers to interaction with neighbors, with Ω_{ij} being the matrix of connections between neurons, $i, j = 1.....n$. Stationary states are: $H_i = +1$ (active) and $H_i = -1$ (passive), that provides the effect of neuron switching on/off under its neighbor's impact. Note that the parameters β referring to the excitation threshold could be modified as the result of learning process.

It should be stressed that the functions performed by the H-type plates depend essentially on the principle of the connection training. Under NCA, two types of training rules are used. The first one that is required for recording corresponds to well-known Hebb's rule [17] of connection amplification, which implies that the strength of connections between excited neurons *increases* as

$$\Omega_{ij}^{\,Hebb}(t) = \frac{\Omega_0}{4 \cdot \tau_\Omega} \cdot \int_0^t \left[H_i(t') + 1\right] \cdot \left[H_j(t') + 1\right] dt' \cdot \zeta(t'), \quad (2)$$

where Ω_0, τ_Ω — training parameters, $\zeta(t)$ is monotonic integrable function to provide the saturation effect.

Another version of the connection-training principle had been proposed in original work of J. Hopfield [20] as a tool for *recognition* of the already learned (stored) images. This version reads:

$$\Omega_{ij}^{\,Hopf}(t) = \Omega_0\{1 - \frac{1}{2\tau_0}\int_0^t [1 - H_i(t')H_j(t')] \cdot \zeta(t')dt'\}, \quad (3)$$

that corresponds to the "redundant cut-off" principle. This means that the "informative" connections between excited neurons are initially strong and do not change in the training process, while irrelevant (waste) connections should die out. This principle corresponds actually not to the *choice* of recording, but rather to the *selection* of trained connections.

It should be stressed that such way of training leads to the fact that this processor could perceive *any* (even new) image as one of the already learned (stored). This results in two effects:

- refinement of the damaged (noisy) image: due to the hard influence of neighbors, the irrelevant neurons would die, while missing ones would be excited;
- there are problems with re-learning of this processor to incorporate new images.

Thus, the necessity and reasons for exploring two versions of the H-type processor are apparent.

2) Symbolic memory:

This concept involves the coding (localization) procedure combined with possibility of further cooperative (Hebbian) interaction. These two functions could be realized by means of the Grossberg-type (***G***) processor [22] with

competitive intra-plate (horizontal) interaction, which works at the first stage for choosing one neuron to be the *symbol* (representer of the certain group of neurons, i.e., the *image*). At the next stage, competitive interaction should be altered to cooperative. The model of processor possessing all these abilities could be written in the form*:*

$$\frac{dG_k(t)}{dt} = \frac{1}{\tau_G}[\{-(\alpha_k - 1)\cdot G_k + \alpha_k \cdot G_k^2 - G_k^3\} - \theta(\Psi_0 - \Psi)\cdot \sum_{l\neq k}^{n} \Gamma_{kl}\cdot G_k \cdot G_l + \vartheta(\Psi - \Psi_0)\cdot \sum_{l\neq k}^{n} \Omega_{kl}\cdot G_l] + Z(t)\cdot \xi(t) \equiv \frac{1}{\tau_G}[\Im_G\{G_k, \alpha_k\} - \theta(\Psi_0 - \Psi)\cdot \sum_{l\neq k}^{n} \Gamma_{kl}\cdot G_k \cdot G_l + \theta(\Psi - \Psi_0)\cdot \sum_{l\neq k}^{n} \Omega_{kl}\cdot G_l] + Z(t) \quad (4)$$

where the variable G_k refers to the state of *k*-th *G*-type neuron, τ^G is activation characteristic time, with its internal dynamics being described by the functional $\Im_G\{G_k,\alpha_k\}$. The term $Z(t)\cdot\xi(t)$ stays for the random component, with $Z(t)$ being the noise amplitude, $0<\xi(t)<1$ is random function. Two step-wise *theta* functions $\vartheta(\Psi-\Psi_0)$, $\vartheta(-\Psi+\Psi_0)$ stop the competitive process and start the cooperation (depending on the argument's sign).

Note that this representation differs from given for the *H*-type neuron since the stable state here are equal to: *G*=1 (active) and *G*=0 (passive). Such choice of representation enables us to account for both, competitive and cooperative interactions depending on the state of inter-plate (so called "vertical") Ψ connections.

The *competitive* connections Γ provide the symbol-choosing procedure that requires mandatory participation of random component (noise), see [11]. The dynamics of connection training is determined by the equation:

$$\frac{d\Gamma_{kl}(t)}{dt} = -\frac{\Gamma_0}{\tau^{\Gamma}}\{G_k \cdot G_l(G_k - G_l)\}, \quad (5)$$

where Γ_0 and τ^{Γ} are training parameters. This training rule provides so-called "localization" reaction, when only one neuron from the activated chain wins the round. So this type of neuroprocessor serves to convert the chain corresponding to the real object (i.e., the "*image*") into single neuron referred further as the "*symbol*" (in other terminology, the "*name"*) of this object.

After the choosing procedure was finished, the inter-plate (vertical) Ψ connections should formed to link the chosen symbol with its image neurons at the previous hierarchy level:

$$\frac{d\Psi_{km}^{R,\sigma}(t)}{dt} = \frac{\Psi_0}{\tau^{\Psi}}\cdot G_k^{R,\sigma}\cdot G_m^{R,(\sigma-1)}, \quad (6)$$

where Ψ_0 and τ^{Ψ} are characteristic parameters of training. Such connections secure the *semantic* content of the chosen symbol; therefore, they are called the "*semantic"* connections. These very connections do realize the Kohonen paradigm "Winner Takes All" (WTA) [38], providing a possibility to decompose the symbol into distributed image.

Note that this processor differs from the standard versions of ACT procedure (see, e.g., [53], [54]) by at least two factors:

- there is no fixed rule for conversion process, it proceeds due to *competitive interaction* between neurons only;
- symbol-formation procedure in the given processor is *unstable*, thus providing uncertainty and "individuality" of the position of chosen symbol.

Let us stress that this very mechanism of the winner-choosing procedure is derived not from the neurocomputing, but from the analysis of choices done within given society, and is known in DTI as the "conventional information struggle" (see [14], [15]). It is typical not only for the human society, but for all living objects as well. This very choice should not be "*the best*" (i.e., the most efficient, or fast, or reliable, as it is typical for neural computing), but should be individual for the given system. Thus, the symbol formation procedure under NCA represents an example of creating the *conventional* information.

After the semantic connections between the chosen symbol and its image were formed up to sufficiently ("black") Ψ_0 value the competitive interaction stops due the presence of step-wise function in (4). Then, the cooperative interaction with neighboring symbols could start that correspond to the last term in (4). The cooperative connections are trained according to the Hebbian principle:

$$\frac{d\Omega^{\sigma}{}_{kl}(t)}{dt} = \frac{\Omega_0}{\tau^{\Omega}}\{G^{R,\sigma}{}_k \cdot G^{R,\sigma}{}_l\}. \quad (7)$$

These connections provide the possibility to form the *generalized* image, i.e., "image-of-symbols", which could get its symbol at the next hierarchy level. Note that this process may be reproduced at each step of the system's evolution. Thus, this processor actually possesses the properties of distributed memory as well.

Note that in our previous works [11], [12] this effect was secured by the mechanism of *parametric modification* of the neuron-symbol, which takes it out from the competitive interaction, simultaneously providing the possibility of cooperative interactions with neighbors. It has been proposed that after the given *G*-neuron got a status of symbol and had formed the inter-plate connections Ψ with his image neurons, it should leave a competitive struggle for the right to be a symbol of another image. This effect could be provided by parametric modification of the neuron-

symbol: $\alpha_k \rightarrow \alpha_k(\{\Psi_{ik}\})$. Actually, both mechanisms, dynamical and parametric ones, could work together.

In any case at the time scale $t >> t^{\Gamma}$, the neuron-symbol stops its competitive interaction with neighbors, but acquires a possibility to participate in the *cooperative* interactions with the other neuron-symbols by the same Hebbian mechanism as *H*-type neurons do. Note that "free" ***G***-neurons (that were failed to become a symbol of any image) could compete only.

Another very important point should be stressed. Encoding (i.e., symbol formation) means as well the *comprehension* of the image information received from outside. The very fact of symbol formation implies that the system had apprehended the given chain of *M* active neurons at the plate ***H*** as a representation of a single real object and had awarded a proper symbol ("name") to it. That is why the inter-plate (vertical) connections between the symbol and its progenitor image neurons are called *semantic* ones.

Let us stress ones more that, the instability of the conversion procedure under NCA results in just *random (free) choice* of the symbol among possible "nominant" neurons. This means that this procedure represents a particular case of generation of *conventional* information – this choice should not be the best (the most efficient), it should be individual. Thus, this process does secure the *individuality* of any (even artificial) cognitive system.

IV. ARCHITECTURE OF COGNITIVE SYSTEM

The architecture of cognitive system has been designed under NCA in the works [11], [12] of Chernavskaya et al. Let us recall briefly main peculiar features.

A. Basic Elements of NCA Architecture

The schematic representation of NCA cognitive architecture is plotted in Fig. 1. This system represents a composition of several neural processors of Hopfield (***H***) and Grossberg (***G***) types, which are composed into hierarchical structure, with σ being the number of hierarchical level. Each processor is represented as a plate populated with *n dynamical* formal neurons described in Section III. The total number of levels (symbolic plates) is neither fixed nor limited since they appear "as required" in course of the system evolution as a response to the operational complexity of the perceptible world.

Each symbol G^{σ} is linked by *semantic* connections $\Psi^{(\sigma-1)}$ and $\Psi^{(\sigma+1)}$ defined in (6) with its "parent" image at the previous level and the "descendant" symbol at the next level σ+1, respectively. Besides, it is linked with its *neighbors* by cooperative connections Ω^{σ} (defined in (7)), which create new (independent) image. Using imagination, one may say that each symbol has its "*legs"* (to rely to the ground) and "*hands"* (to reach the ceiling). Such "pyramid" is replicated at every level of hierarchy, thus forming the fractal-type multi-level structure.

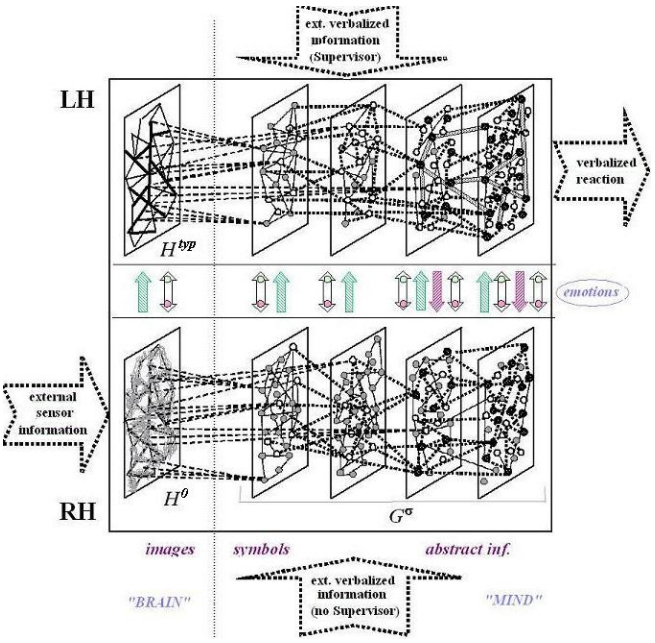

Figure 1. Schematic representation of NCA cognitive **a**rchitecture.

According to DTI principles, the system is divided into two coupled subsystems, the right hemi-system (**RH**) and the left hemi-system (**LH**). These terms were chosen to correlate these subsystems with cerebral hemispheres, with the cross-subsystem connections Λ(t) being an analogue to the *corpus callosum.* These connections should provide the interaction ("dialog") between the subsystems ("up-down" arrows in Fig. 1). One subsystem (**RH**) is responsible for *learning* and processing *new* information; the other one (**LH**) is dealing with the *well-known* information. This functional specialization coincides completely with that proposed (from the "mind" viewpoint) by Goldberg [19], that represents a pleasant surprise as well as an indirect validation of our approach. Under NCA we can also *reveal its mechanism* from the "brain" viewpoint. It is secured by three factors:

- the presence of random component (noise) in **RH** provides the conditions for generation of information, i.e., *free choice* of the version of recording new information;
- different connection-training principles in the different subsystems: the Hebb's principle of active connection amplification [17] in **RH**, and the Hopfield's principle of the "redundant cut-off" [20] in **LH**;
- the "connection-blackening" principle of self-organization, which implies that strong enough ("black") images in **RH** are replicated in **LH**. Hence, **RH** acts as a Supervisor for **LH**.

Let us consider the connection-blackening principle in more details by analyzing the elementary act of system's evolution.

B. *Elementary Learning Act: "Connection-Blackening" Principle*

The elementary act of cognitive process realization (in particular, learning) should involve implementation of the functions of recording, storing and coding the image of new object.

The functions of recording and storing "raw" images could be implemented by means of two *H*-type cross-linked processors (see Fig. 2a), with the connection-training rules being *different* on those plates (Fig. 2b). One of them (called $\boldsymbol{H^0}$) should be trained by Hebbian mechanism, while the other one (called $\boldsymbol{H^{typ}}$) — according to the original Hopfield principle "redundant cut-off". They are correlated by the value of well-trained connections Ω_0 (see Fig. 2b).

Primary ("raw") images are recorded at the plate $\boldsymbol{H^0}$ by Hebbian-trained connections, with their strength being vary from weak ("grey") to strong ("black") state. When the strength of trained connections achieves the "black" value Ω_0, the "black" image should be transferred by direct (one-to-one) inter-plate connections and replicated at the *typical image* plate $\boldsymbol{H^{typ}}$ for storing. This procedure corresponds to the implementation of so called "connection blackening" principle.

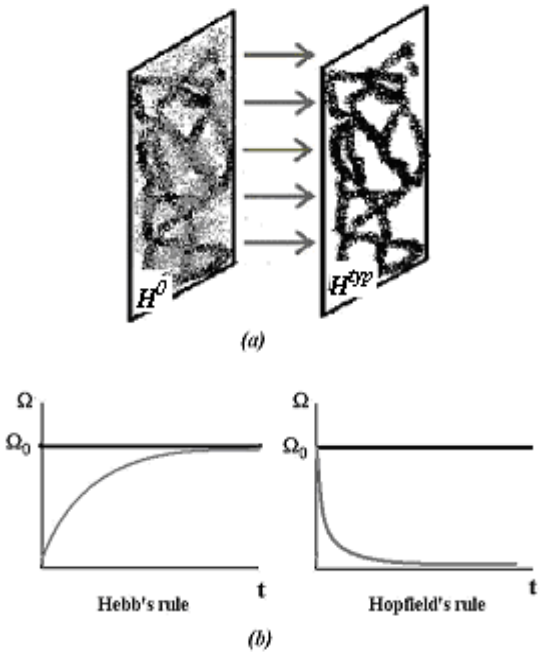

Figure 2. Schematic representation of recording and memorizing process (a) and (b) time dependence of corresponding intra-plate (horizontal) connection strength Ω(t).

The combination of this process with the encoding procedure provides the "*elementary act*" of the system's formation and is presented in Fig. 3. This process again corresponds to the self-organization principle of "connection blackening" and proceeds in three steps:

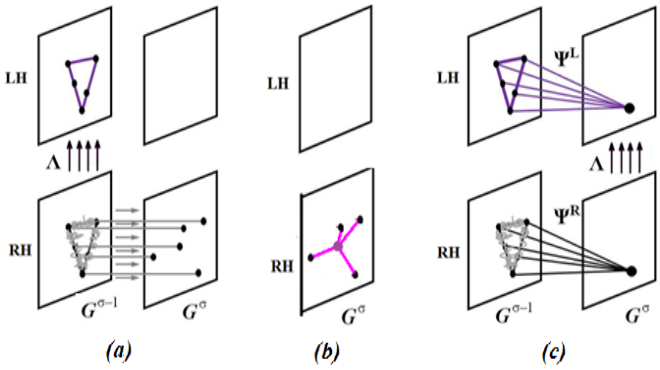

Figure 3. The elementary act of learning.

a) The First Step: an image formed at the previous-level (σ–1) in **RH**, after its cooperative connections Ω^R become strong ("black") enough, is delivered by the direct (one-to-one) inter-plate ("vertical") connections to the next-level plate $\mathbf{G^{\sigma}}$ and, simultaneously, by the inter-subsystem connections Λ to the same level plate $\mathbf{G^{\sigma-1}}$ in **LH** (see Fig. 3a); **LH** level is free.

b) The Second step: NCA conversion procedure image-into-symbol occurs at the next-level plate $\mathbf{G^{\sigma}}$ in **RH** (Fig. 3b); **LH** level is free.

c) The Third (Final) Step: New symbol is formed together with its semantic (one-to-many) inter-plate connections Ψ^R and is replicated at the same level in **LH**, where vertical connections Ψ^L are forming according to Hopfield-type rule. Here again, the "connection blackening" principle for Ψ^R connections controls the symbol-formation process (Fig. 3c).

This process could be repeated at each level of hierarchy thus generating a multi-fractal structure.

It is important to stress that the raw images in **RH** with relatively weak ("grey") connections (those that didn't achieve the level typical for **LH**) are neither transferred to the next level in **RH**, nor replicated in **LH**. They remain only at the given level and not acquire their symbol at the next level. Thus, they represent latent (hidden) information, which is "auxiliary" for the given system.

C. *Specialization of Various Hierarchical Levels*

Let us discuss the roles of different hierarchy levels and their correspondence to the cerebral functional areas.

1) Hierarchy-Level Specialization: The whole system represents complex multi-level block-hierarchical construction that does evolve by itself (in Fig. 1 — from the left to the right) due to the self-organization principle of "connection blackening". This implies that at each level, the elementary act presented in Fig. 3 is repeated. New levels (symbol layers) appear "as required", i.e., after a new image was formed at the previous level. In physics, there is special term "scaling" for such principle of organization and the whole structure is called a fractal.

The lowest level σ = 0 is represented by the *H*-type plates containing the *image* information. The plate $\boldsymbol{H^0}$ in

RH carries the *whole* image information received by the given system by means of the "sense organs", i.e., from the receptors. The intra-plate (horizontal) connections vary from weak ("grey") up to strong ("black") ones. Note that the images recorded by "grey" (rather weak) connections, according to the connection blackening principle described above, are neither delivered to the next level, nor replicated in **LH.** They are stored at H^0 only, thus representing some vague (fuzzy) information. That is why the plate H^0 hereinafter is referred to as the "fuzzy set". This plate is responsible for recording new sensor images.

The plate H^{typ} in **LH** contains the information selected for storing (memorization). This plate is "filling up" in course of learning (with the role of Supervisor being played by the plate H^0**)** with those images that are recorded by sufficiently "black" connections ("up" green arrow in Fig. 1). These images are referred to as *typical* ones. This plate does play the main role in recognition of already learned objects; in some sense, it is a classifier.

The next level $\sigma = 1$ is occupied by the *symbols of typical images*, which are formed in **RH.** These symbols do carry a semantic content, that is, a *comprehension* of the fact that the given chain of active neurons represents one real object. Semantic content (sense) of such symbol consists in its *decomposition* (by means of semantic inter-plate connections Ψ) into its image corresponding to this very real object. Only after formation of sufficiently "black" connections Ψ^R, this symbol could be replicated in **LH**.

At the same very level, the process of *primary verbalization* starts. This implies that there occur the *internal words* as the *names* of already learned objects. These names occur in **RH**, i.e., they are chosen *arbitrary* and individually thus are understandable for a given system only. If simultaneously **LH** receive an external information (from external Supervisor, see top external arrow in Fig. 1) on conventional name for this object, the 'internal" name would be replaced (after certain conflict) by the conventional one (by means of inverse training **LH→RH**, see "down" purple arrow in the middle part in Fig. 1). Such process, that is similar to the process of children speech trials, was considered and discussed in [14], [15].

At the same level in **RH**, the symbols could cooperate and create the *generalized images* (image-of-symbols), which acquire their own symbols at the next level $\sigma+1$. These images are rather primitive, since they correspond to concrete real objects. However, even at this level, a Poet could create, using primitive words, a pronounced pattern ("*night, street, lamp, drugstore…*" as in a famous Alexander Block's poetry).

At the next levels $\sigma>1$, this process is repeated with increasing degree of "abstraction" of created images. This implies that new generalized images could hardly be related to any real object and explained at the image level.

At the higher levels of hierarchy $\sigma>>1$, the *abstract information* emerges, that is, the infrastructure of symbols and their connections, which are not mediated by "raw" images, i.e., the neuron-progenitors of *H*-type plates. Here, the *concept symbols* arise, that could not be related to any concrete pattern (e.g., *conscience, infinity, beauty, consciousness, love,* etc.). This information appears in the already well-trained system as a result of interactions of all the plates (not "perceptible", but "deduced" knowledge). This very information could be completely *verbalized*, i.e., expressed in the symbolic form (with relevant grammar and syntax) by means of *conventional language* of a given society. These very higher levels provide a possibility of communication with similar systems. This implies a possibility to propagate personal conventional information ("to explain by words") and understand semantic content of external symbolic (verbal) information. Besides, at such level **LH** obtains a possibility to receive new information not only from **RH**, but also from outside, in symbolic form, from external Supervisor. In psychology, such knowledge is called "semantic", in distinguish to "episodic" one that the system (**RH**) obtains in process of acquiring its individual experience. This knowledge could appear to be active only after incorporation into the existing architecture due to **LH→ RH** connections ("down" purple arrow at the right part of Fig. 1). Note that **RH** itself can get the symbolic verbalized information from outside, without Supervisor (bottom external arrow in Fig. 1), and this information is processing just as internal one, i.e., by forming the Hebbian connections between different external symbolic images.

Thus, the system as a whole does *grow up* from the lower *image* information levels, over *semantic* information (understandable for a given individual system only), to the higher levels of *abstract* information, which could be verbalized and *propagated* (understood) within the given society. At every stage of new level formation, the same process is repeated. New connections are forming in **RH** up to the "black" state, and after that, the new-formed symbol is transferred to **LH**. In this process, certain part of information (*inessential* details recorded by "grey" connections) appears to be lost. Speaking more exactly, it is not delivered to the next level, but is stored at the previous one as *auxiliary* or *latent* information specific for a given individual system.

Note that the label "emotions" in Fig. 1 refers neither to **RH** nor to **LH.** Below, it will be shown that emotions are directly related to switching the cross-subsystem connections Λ ("up-down" arrows in Fig. 1) providing the "dialog" between subsystems. The color of arrows reflects emotional "valence" (green for positive and rose for negative ones).

2) Corresponendence with Cerebral Cortex Areas: Let us point out that the geometry of the NCA architecture corresponds to the functional areas of the human cerebral neocortex (see Fig. 4). The neocortex could be (conventionally) divided into areas ("lobes"), which are responsible for the vision (occipital lobe), motor activity (parietal lobe), auditory activity (temporal lobe), abstract thinking (frontal lobes), etc. Temporal lobes embraces Wernicke's and Broca's areas that are responsible, respectively, for language hearing (word perception) and reproducing (word production), but not for the speech itself.

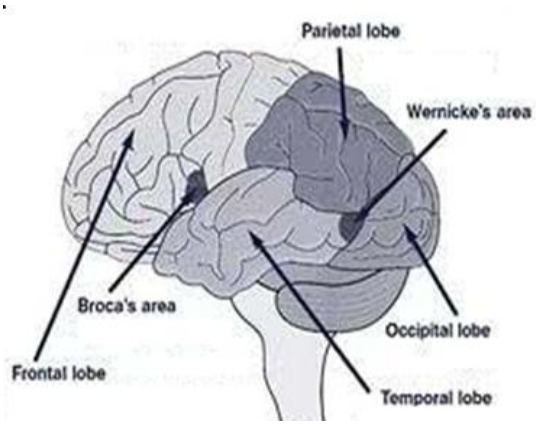

Figure 4. Map of the functional areas of human cerebral cortex (extracted from [55])

The speech function, i.e., coherent and sensible transmission of information, relates to the frontal lobe that is associated with abstract thinking.

Note that similar allocation of functional levels is realized in the NCA scheme: low levels (σ=0) provides *images*, i.e., visual patterns; middle levels (σ>1) contain symbol-words, that is, elements of *language.*

The correspondence between the *abstract information* in the NCA architecture (σ>>1) and the "abstract thinking" typical for the frontal lobes, is obvious. Thus, the map in Fig. 4 actually corresponds to the mirror reflection of the scheme in Fig. 1.

D. Interpreting the Concepts of Intuition, Sub-consciousness, Consciousness, and Logic

Now, let us turn to interpretation and revealing the mechanisms of specific human features of cognitive process, — namely, intuition, logic, sub-consciousness, etc.

If the *intuition* is treated as occasional, spontaneous, unreasoned solution, or, following Immanuel Kant [56], "the direct discretion of the truth" without any reasons and proofs – then, it apparently emerges from **RH** (more exactly, from the noise in **RH**). Typical feature of intuition consists in unconscious way of getting the result.

Treating the *logic* as all the cause-and-effect unbroken chains (causal relationships), one could infer that all the processes in **LH** are related. In this sense, the inference of our early paper [57] (where there was no symbolic structure) remains still valid. At this level, the inference of [50] seems valid also.

However, these concepts could be considered in more detail. Thus, the *logical thinking*, according to [28], is defined as "*correct provable reasoning*". This definition immediately leads to the inference that only verbalized reasoning (thereby, conclusive and commonly understandable) is related. At that, the term "correct" implies that these reasons should be based on the *conventional axioms.* Then, between the "pure logic" and "pure intuition" there should be a place for some other, intermediate, thinking algorithms.

Similar reasons concern the concepts of *consciousness* and *sub-consciousness*. Defying the consciousness as "the state of being aware of and responsive to one's surroundings" [28], we infer that it could emerge after verbalization only.

The sub-consciousness is defined as "…aggregate of processes lacking the subjective control" [28]. This implies that it should be based on the randomly stored information, that had not acquired any symbol and thus, could not be activated from outside by means of symbols (i.e., words).

Keeping in mind previous reasons, we can interpret the notions of intuition, logic, and (sub-)consciousness under NCA.

The architecture described above has large number ($\boldsymbol{N}$>>1) of levels. The lower levels contain *auxiliary* or *hidden individual* information for a given system, the "thing in itself". Only verbalized information that occurs at higher levels of hierarchy could be comprehended in a common sense (not individually). Then, we can try to answer the question "How the brain makes a thought?" Since a speech represents a *consecutive set of symbols*, this is the very tool to form (separate) a pattern called a "thought" from all the variety of the brain-activity patterns. There exists a picturesque formula "the language is a means for our brain to speak with us". Thereby the *consciousness* could be defined as the system's ability to draw up the cognitive activity into consecutive content set by means of a speech. Here, the main role is played by **LH**.

As it was shown above, a part of information appears to be lost at any transition from previous level to the next one. More exactly, it converts into form of "*latent*" (auxiliary), or "*hidden*" information for a given system. Let us consider this in more details.

The innermost level of latent information is represented by weak ("grey") connections at the *fuzzy* set, i.e., the image plate $\boldsymbol{H^0}$. Its role consists in storing the "occasional" (i.e., "randomly collected") information that could appear to be important some time later. This information is transferred neither to **LH** nor to the level $\boldsymbol{G^1}$, thus, could not be associated with any symbol. This means that it remains *not comprehended* and not *controlled* by the system, i.e., just what has been defined as the "sub-consciousness". Such ("grey") chains could be activated only due to the noise, by chance ("to see suddenly by internal view"), that could be interpreted as the "aha *moment*" (see, e.g., [26], [58]).

At the transition from semantic information to verbalized one, there remain a lot of symbols that are not associated with any standard word. This implies certain "pictures" that could be described only by means of decomposition, i.e., one internal symbol can be described by several standard words. Verbalization of this information requires not an insight, but assortment of necessary words. This is always possible, but not always simple. Using the terms of recognition theory, this process could be called "formalizing the expert knowledge".

Thus, the latent information is disposed at various levels of depth, and this fact does control the efforts for extracting it up to the consciousness level. It seems natural to interpret the inferences based on the latent information, as intuitive

thinking (*insight*). It is worth noting that in the proposed scheme, the majority of latent information is actually concentrated inside **RH**.

Logical thinking could be specified as "…unbroken sequential thoughts" [27], as well as "…operating by verbalized (abstract) concepts and their connections" [28]. This process is typical for **LH** at higher hierarchy levels.

An abstract information as itself has its own levels and infrastructure, which emerges gradually, in course of system's evolution (for human beings, this implies "with years"). This developed infrastructure that combines higher levels of **RH** and **LH** could be associated with the *wisdom*. This implies that the wisdom is broader than logic.

Specific features of the "latent" elements become rather pronounced in the process of solving the problems related to fixing the similarity/difference of the objects. These problems are solved automatically, at the image levels. The similarity is emphasized by shared neurons, while the difference is specified by diverse ones, and the system *does know* it. However, this knowledge could not be *comprehended* until those common/diverse neurons were not associated with combinations of internal symbols. Then, the *auxiliary-image* knowledge ("feeling") could be converted into *semantic* one. Further verbalization of this knowledge implies ascertainment of the connections between internal symbols and the words. The obtained result is valuable for a given system (individual), but could appear to be fault objectively, since the mode of recording the image information is individual as well. The obtained solution is *intuitive*, since it is based on the recorded experience, i.e., the individual "worldview pattern". This solution should not be proved (the system itself does not need any proof since it just knows that it is so). However, being verbalized, this solution could be explained to others and argued. If the arguments fit the conventional axioms, it would be a proof of its truth. Actually, the method of "converting the intuitive expert knowledge into logic one" is presented aforesaid.

E. Master Equations: Mathematics & Phylosophy

The mathematical foundations for the architecture presented in Fig. 1 were discussed in details in [11]. Let us recall the key points and present the mathematical basis in generalized form:

$$\frac{dH_i^0(t)}{dt} = \frac{1}{\tau_i^H}[\Im_H\{H,\beta_i(G^R{}_{\{i\}})\} + \sum_{i\neq j}^{n}\Omega_{ij}^{Hebb}H_j^0 + \sum_k \Psi_{ik}G_k^{R,1} - \Lambda(t)\cdot H_i^{typ}] + Z(t)\xi_i(t) \quad , \quad (8)$$

$$\frac{dH_i^{typ}(t)}{dt} = \frac{1}{\tau_i^H}[\Im_H\{H,\beta_i(G^L{}_{\{i\}})\} + \sum_{i\neq j}^{n}\Omega_{ij}^{Hopf}\cdot H_j^{typ} + \sum_k \Psi_{ik}\cdot G_k^{L,1} + \Lambda(t)\cdot H_i^0] \quad , \quad (9)$$

……

$$\frac{dG_k^{R,\sigma}}{dt} = \frac{1}{\tau_G}[\Im_G\{G_k,\alpha^\sigma{}_k(\{\Psi_{ik}^{R,(\sigma-1)}\},G^{\sigma+\nu})\} + \hat{Y}\{G_k^{R,\sigma},G_l^{R,(\sigma+\nu)}\} - \Lambda(t)\cdot G_k^{L,\sigma}] + Z(t)\cdot\xi(t) \quad , \quad (10)$$

$$\frac{dG_k^{L,\sigma}}{dt} = \frac{1}{\tau_G}[\Im_G\{G_k,\alpha^\sigma{}_k(\{\Psi_{ik}^{L,(\sigma-1)}\},G^{L,(\sigma+\nu)})\} + \hat{Y}\{G_k^{L,\sigma},G_l^{L,(\sigma+\nu)}\} + \Lambda(t)\cdot G_k^{R,\sigma}] \quad . \quad (11)$$

Here, variables H_i and G_k refer to purely "rational" components that are associated with neocortex, various 'τ' parameters stay for characteristic times. The term $Z(t)\xi(t)$ corresponds to the random (stochastic) component (noise) which is presented in the subsystem **RH** only*; Z(t)* is the noise amplitude. The functionals $\Im_H\{H,\beta\}$ and $\Im_G\{G,\alpha\}$ describe the internal dynamics of corresponding neurons; the functionals $Y^R\{G^\sigma,G^{\sigma+\nu}\}$ and $Y^L\{G^\sigma,G^{\sigma+\nu}\}$ in the equations for symbolic plates describe the horizontal and vertical interactions of symbols (see [11] for details); $\Lambda(t)$ specifies the cross-subsystem connections.

Let us present several remarks on the meaning of certain terms.

1) "Brain vs. Mind" Border:

First two equations relate to the lowest (zero) level of hierarchy, while the others (***G***) variables describe σ=1,…N symbolic levels. Note that the *dotted line* after two first equations indicates the analogy with the dotted line in Fig. 1. This line symbolizes the *virtual border* between the Brain and the Mind. Indeed, the ***H***-plates (zero-level of the hierarchy) containing only the "raw" images, serve to represent the sensible information received from the organs of sense. This information is (roughly speaking) objective, so this level belongs (roughly speaking) to the Brain.

The level σ=1, that is, the level of the typical-image symbols, already belongs to the Mind, since any symbol represents not objective, but *conventional*, i.e., *subjective* and *individual* (for a given system) information. The same is true even more for all other hierarchy levels, up to the highest level associated with the abstract information. Thus, we come to

Philosophical Inference #1: The "bridge" between the "Brain" and the "Mind" is made of semantic connections between symbols and their images, i.e., by conventional

(individual) information generated by the neuron ensemble itself.

2) Reflection of a single-neuron history:

The functionals $\Im_H\{H,\beta\}$ and $\Im_G\{G,\alpha\}$ defined by (1) and (4), respectively, describe the internal dynamics of the corresponding *dynamical formal neurons*. This very representation provides the possibility to describe the *parametric mutations* of the "trained" neurons, — i.e., those neurons that actually participated in creation of images and symbols forming the architecture as a whole. This effect corresponds to experimental evidences from [18].

One of the parametric-modification mechanisms consists in the influence of high-level symbols on the corresponding image neurons: $\beta_i \to \beta_i\{G^{\sigma}_{\{i\}}\}$. First of all, this refers to so called *symbol of class*, that is, the symbol, which was induced not by the image of certain object, but by a set of *common attributes* of certain class of objects. Excitation of such symbol could not excite all "referring" images, but switches them into the "standby mode" by lowering the activation threshold of common image neurons β_i. Thus, these images acquire the *right of priority* for activation, i.e., an *attention.*

All these arguments refer as well to the parameters α^{σ}_k of symbolic neurons. The *k*-th neuron at the plate $\boldsymbol{G^{\sigma}}$, being a member of new "generalized" image, plays the role of the *image* neuron for all the higher-level symbols $G^{\sigma+\nu}{}_{\{k\}}$ that it is related, thereby its parameter should be modified as: $\alpha^{\sigma}_k \to \alpha^{\sigma}_k\{G^{\sigma+\nu}{}_{\{k\}}\}$. Besides, as it was considered above, the neuron-symbol should be modified parametrically after its semantic content (i.e., the inter-plate connections $\Psi^{(\sigma-1)}{}_{ik}$ with its image) was formed: $\alpha^{\sigma}_k \to \alpha^{\sigma}_k(\{\Psi^{(\sigma-1)}{}_{ik}\})$. This modification takes the neuron out from the *competitive* interactions and turns on the *cooperative* ones. This factor secures complex multi-level interactions of the neuron-symbols and leaves "off screen" those *G*-neurons that failed to become a symbol.

Thereby, complete modification of a *G*-neuron reflecting the "history" of his relations with other neurons (his "skill") could be presented in the form: $\alpha^{\sigma}_k \to \alpha^{\sigma}_k(\{\Psi^{(\sigma-1)}{}_{ik}\}, G^{\sigma+\nu}{}_{\{k\}})$.

Thus, the model of dynamical formal neuron enables us not only to reproduce the fact of mutation of the "trained" neurons observed in [18], but also to specify and distinguish concrete modifications associated with different "skills".

Philosophical inference #2: The account for the neuron internal structure enables us to reproduce the effect of mutation of the neurons participated in certain "skill" acquirement. This provides the interpretation for the effect of "neuron memory" concentrated not in the inter-neuron connections, but inside the neurons themselves.

3) What is the tool for switching the subsystem activity?

The variable $\Lambda(t)$ controls the dialog between two subsystems. This is the only variable presenting in each equation, thus 'sewing' all the components together. Therefore, it deserves special discussion. These connections should not be trained, but should provide *switching* the subsystem activity in course of the problem solving. Here, the connections $\Lambda^{R\to L}$ activating **LH** are treated as positive $\Lambda^{R\to L} = +\Lambda_0$, and vice versa, connections $\Lambda^{L\to R}$ activating **RH** are treated as negative ones $\Lambda^{L\to R} = -\Lambda_0$. All the processes requiring the generation of new information, namely — forming either new image, or new symbol — are to proceed in **RH** with necessary noise participation. Then, the result of this process should be transferred to **LH** by direct cross-subsystem connections: $+\Lambda_0$. The reverse connections $-\Lambda_0$ are switching on in the already trained system, when an incoming external information appears to be unknown, i.e., *new*. Then, the system should pass over the re-training stage by means of **RH.** Let us stress that the mechanism of the $\Lambda(t)$ switching is not specified in (8) – (11) yet; it will be considered in the next Section.

Note that this system of equations is not complete in mathematical sense (as it was also in [11]), since not all the variables are determined via their mutual interactions. Namely, Z(t) was considered as a model parameter, and the mechanism of $\Lambda(t)$ switching is not clear. Since the considered cognitive architecture is in a good agreement with functional areas of *neocortex* (not subcortical structures), we come to

Philosophical Inference #3: Proper system of equations that describes the whole cognitive process could be completed only after taking into account the participation of *emotions.*

V. The Role and Place of Emotions

The incorporation of emotions and rational thinking into cognitive system represents really the challenge, since we need to ride over the explanatory gap between "Brain" and "Mind". Under NCA, this implies that two different "tools" are required, the one relating to the "Brain" structures, and the other one expressed in the "Mind" terms. Then, mutual influence of these "tools" could provide *integral* representation of emotions in the cognitive process.

From the evolutionary point of view (see, e.g., [30]), emotions represent far more ancient mechanism of the analysis of environment, than rational reasoning. Therefore, the sources of emotional bursts relate to so called "old cerebellum", — i.e., certain sub-cortical structures like *thalamus*, *basal ganglia*, *amygdale*, *substance negro*, etc. (see [7], [30]). Then, the production of these very structures could be considered as the required "Brain tool" for emotion representation.

From the other hand, the rational reasoning as rather "young" (evolutionary) ability relates to cerebral *neocortex*. Thus, the required "Mind tool" should relate also to this very structure.

Emotions provide a *synthetic* (integral) reaction that appears before the analysis of concrete reasons and motives. For humans, the specification of "*emotio*" and "*ratio*" becomes meaningful after formation of the common *language* (that is, the developed system of conventional symbols) within a certain community (see, e.g., [45]). Let us point out that any language-delivered information (speech) represents a *successive time set* of symbols. Hence, the reasoning, or rational thinking, represents a *consecutive* method of information processing. Therefore, it seems reasonable to assume that not-rational or emotional

reactions correspond to the *parallel* information processing. Recalling that these functions are attributed to the left and right hemispheres, respectively (see [25], [50]), one may come to a big *temptation* to infer that rational and not-rational (emotional) thinking correspond to **LH** and **RH**, respectively. Below, it is shown that that all these arguments actually are related to the problem, but realization of this program calls for more accurate consideration.

A. The Problem of Emotion Formalization

In order to formalize the above arguments, let us consider the approaches to emotion classification.

In psychology, the self-appraisal (emotion) is ordinarily associated with achieving a certain *goal*. Commonly, emotions are divided into positive and negative ones, with increasing probability of the goal attainment leads to positive emotions, and vice-versa. Furthermore, it is known that any *new* (*unexpected*) thing/situation calls for *negative* emotions (see, e.g., [19]), since it requires additional efforts to hit the new goal (in the given case, to adapt to unexpected situation). Hence, to the first approximation emotions could be divided into positive and negative ones.

From the neurophysiology viewpoint, emotions are controlled by concentration and composition of the neurotransmitters inside the organism [7], [25]. All the exciting variety of known neurotransmitters (more than 4000 known species) can be sorted into two groups: the *stimulants* (like *adrenalin*, *caffeine*, etc.) and the *inhibitors* (*opiates, endorphins*, etc.). Note that this fact indicates indirectly that the binary emotion classification — positive vs. negative ones — seems bearable despite its primitiveness. However, there is no direct correspondence between positive self-appraisal and the excess of inhibitors or stimulants, the problem is more intriguing.

Anyway, the simplest "Brain tool" to represent the emotions is rather apparent: it is the *effective* (aggregated) *composition* of neurotransmitters $\mu(t)$ representing the *difference between the stimulants and inhibitors.*

According to DTI, emotions could be divided into two types: *impulsive* (impelling the generation of information) and *fixing* (effective for reception). Since the generating process requires the noise, it seems natural to associate impulsive emotions (*anxiety, nervousness*) with the *growth of noise amplitude Z(t).* Vice-versa, fixing emotions could be associated with *decreasing* noise amplitude (*relief, delight*). By defining the goal of the living organism as the maintenance of *homeostasis*, (i.e., calm, undisturbed, stable state), one may infer that, speaking very roughly, this classification could correlate with negative and positive emotions, respectively.

Thus, we may infer that it is the noise amplitude *Z(t)* (relating actually to the *neocortex*) that could be treated as the required "Mind tool" for accounting emotions.

B. Main Hypotheses on Emotion Representation in AI

We propose the following hypothesis on the nature of emotions: *The random component (noise) in artificial systems does correspond to the emotional background of living systems, as well as free (random) choice imitates the human emotional choice.*

This concept gives immediately *three* tools directly connected with emotions, and all of them are individual for any given artificial system:

Z_0 — stationary-state background, i.e., the value that characterizes the state "at rest";

$\Delta Z(t) = Z(t) - Z_0$ is the excess of the noise level over the background, which reflects the *measure* of cognitive activity;

dZ/dt — is the time derivative of the noise amplitude, which apparently is the most promising candidate to the analogue to emotional reaction of human being. The absolute value of derivative dZ/dt corresponds to the *degree* of emotional manifestation: drastic change of noise amplitude imitates either *panic* ($dZ/dt>0$), or *euphoria* ($dZ/dt<0$), and so on.

Various combinations of these values reveal a wide field for speculations and interpretations. For example, the calibrated value Z_0 could serve as the indicator of *individual temperament.* The states with $Z(t) < Z_0$ could be interpreted as *depression*, etc. These parameters could be applied to construct artificial cognitive systems (*robots*) of various "psychology" types.

The influence of the "Brain" component should be accounted by linking the value of dZ/dt with an *aggregated* variable $\mu(t)$ that represents the *effective* composition of neural transmitters. In an artificial cognitive system and AI, an additional (artificial) variable $\mu(t)$ should be introduced as an external factor to control the "emotional" state of the system.

Thus, the Main Hypothesis results in the following set of basic hypotheses:

- ***Hypothesis #1*:** The impact of neurotransmitters should be described by the system of equations linking the noise amplitude *Z(t)* with the aggregated variable $\mu(t)$ that corresponds to the effective composition of neural transmitters (the difference between *stimulants* and *inhibitors*).
- ***Hypothesis #2*:** The *apprehended* emotional reaction of human beings could be described as the *time derivative* of the noise amplitude *dZ(t)/dt.*

Note, that this value could be either positive or negative that could be (very roughly) related to negative and positive emotions, respectively. The absolute value of derivative corresponds to the *degree* of emotional manifestation and can take *any* values to describe various emotional shades.

- ***Hypothesis* #3:** The same derivative should control the "dialog" between subsystems: increasing *Z(t)* (*negative* emotions) corresponds to activation of **RH**, while decreasing Z(t) (*positive* emotions) switches on the **LH** activity.

Basing on these hypotheses, we can write the system of equations describing mutual interaction of the variables $\mu(t)$ and Z(t) in course of cognitive process in the form:

$$\frac{dZ(t)}{dt} = \frac{1}{\tau^Z} \cdot [a_{Z\mu} \cdot \mu + a_{ZZ} \cdot (Z - Z_0) + F_Z(\mu, Z) + X\{\mu, G_k^{R,o}\} + \{\chi \cdot (D - \omega \cdot dD/dt) - \eta \cdot \delta(t - t_{D=0})\}], \quad (12)$$

$$\frac{d\mu}{dt} = \frac{1}{\tau^\mu} \cdot [a_{\mu\mu} \cdot \mu + a_{\mu Z} \cdot (Z - Z_0) + F_\mu(\mu, Z)], \quad (13)$$

$$\Lambda(t) = -\Lambda_0 \cdot th\left(\gamma \cdot \frac{dZ}{dt}\right), \quad (14)$$

where a, χ, η, τ, ω, and γ are model parameters, the functional $X\{\mu, G_k^{R,\sigma}\}$ refers to the process of new symbol formation (which decreases $Z(t)$ value, see details in [12]). The linear in Z and μ part in (12), (13) provides the system's homeostasis: stationary stable state corresponds to $\{Z=Z_0, \mu=0\}$. The functions $F_Z(\mu,Z)$ in (12) and $F_\mu(\mu,Z)$ in (13) are written to account for possible nonlinear effects, which may arise from mutual influence of "emotional" (neurophysiology) and "rational" (referring to the neocortex ensemble) variables (see below).

The last term in (12) refers to processing the incoming information. The term D stays for the *discrepancy* between the *incoming* and *internal* (stored) information that provokes $Z(t)$ *increasing.* This very situation refers to the "effect of surprise", which evokes human's negative emotions. Vise versa, finding the solution to the problem (D=0) results in momentary decrease of $Z(t)$, that corresponds to positive emotional splash. Thus, the model seems quite reasonable.

Finally, the hypothesis #3 results in (14), where Λ_0 being the characteristic value of the cross-subsystem connections; γ is the model parameter, which specifies the Λ dynamics. Note that *hyperbolic tangent* function in (14) provides the step-wise behavior at $\gamma>>1$. This implies that $\Lambda = \Lambda_0 = \Lambda^{R\to L}$ at $dZ(t)/dt<0$ and $\Lambda = -\Lambda_0 = \Lambda^{L\to R}$ at $dZ(t)/dt>0$, with Λ being zero at $dZ(t)/dt=0$. Small/moderate variations of dZ/dt around zero provide corresponding oscillations of $\Lambda(t)$ that represent permanent (normal) "dialog" between subsystems. Besides, the solution to standard problems can be found in **LH** only and commonly does not provide any emotional reaction: $\Lambda \sim dZ/dt = 0$ (any inter-subsystem connections are not activated). Hence, this equation fits completely our previous psychological considerations.

Thus, the system of equations (8) – (14) appears to be fully complete since all the variables are defined via their mutual interactions. Let us stress that linking the cross-subsystem connections $\Lambda(t)$ with the emotional variable $dZ(t)/dt$ gives quite original and necessary mechanism to control the subsystem activity and provides desired tool for realization of an artificial two-subsystem schemes (robots).

C. Application of the Model to the Stress/Shock Effect

Let us consider an example of applying this model to reproduce certain observable effect. The effect of "stress and shock", that occurs when people find themselves in a stressful situation, was investigated for several years by the group of neurophysiologists [59]. Two specific characteristics of electrocardiogram were measured, one of them being an appraisal of vegetative imbalance, another one being the measure of heart-rate variability. It was observed that under small or moderate external impact, people gradually calm down after several oscillations of measured characteristics. But in the case of strong impact, initial excitation changes for *depression* and only after sufficiently long time the person can return to ordinary (regular) reactions. This type of behavior was identified as "*stress"*. Moreover, there was detected the regime called a "*shock*": the probationer, after too strong initial excitation, falls down to *deep depression* (*stupor* or *coma)*, and cannot relax independently, without medical assistance. In the latter case, the vegetative balance is controlled by the *opiates* only (pronounced inhibitors), with the variability index comes to zero. It is worse noting that the levels of initial excitation resulting in "irregular" regimes of behavior were detected to be individual. All these regimes could be reproduced in the proposed model by choosing an appropriate parameter set.

The first attempt to describe these effects was done in [12], where *two different* sets of parameters had been used to reproduce the "normal\stress" and "shock" regimes respectively. This means that, the transition between the stress and shock states was treated as *parametric* modification of the system. Alternative version of this model (different choice of parameters) is presented in this article. It enables us to reproduce *all the regimes* within *single* combination of parameters, by varying the initial conditions. Besides, modern description of the stress-to-shock transition seems to be more interesting and relevant (see below).

In Fig. 5, the phase portrait for the model (12) – (13) is presented, where the parameters are chosen to provide the *N*-shape isoclinic curve $dZ/dt = 0$ with just *two stable* statio-

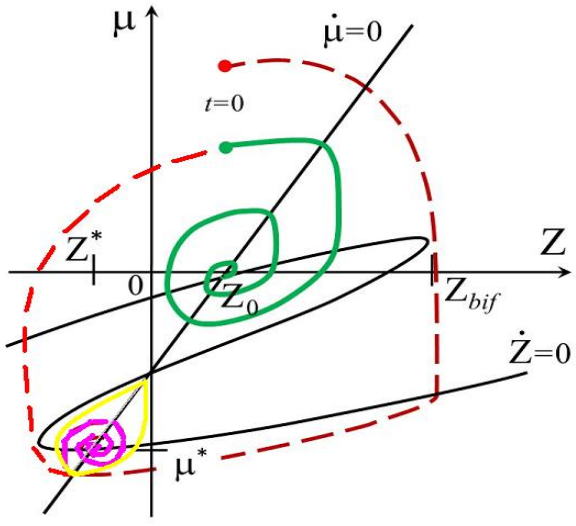

Figure 5. Model phase portrait in terms of "noise amplitude $Z(t)$ vs. an aggregated neurotransmitter composition $\mu(t)$".

nary states. The normal stationary state {Z=Z_0, μ=0} corresponds to homeostasis. The second one {Z=Z*, μ=μ*} corresponds to *abnormal* state (pathology), where the noise is deeply suppressed (Z*<0), and the neurotransmitter imbalance is shifted to deep inhibitor region (μ*<<0). This state just corresponds to that of the "shock" — this implies deep depression with possible transition into a *coma*. Both stationary states represent *stable focuses*.

Normally, the dynamical regime represents *damping oscillations* around the homeostasis point {Z_0,0}. Initial excitation μ(t=0) (that imitates an external impact) provokes growth of *Z* supplied by the following decrease of μ down to negative values, which then changes for decreasing Z with μ growth, and so on. Thus, the values of *Z(t)* and μ(t) gradually (over several cycles) trend to their stable points (solid green curve). But if the trajectory, starting from somewhat larger initial value μ(0), would pass beyond some *bifurcation* value Z_{bif}, the dynamical regime changes (dashed red curve). The trajectory falls down to negative μ (inhibitor) zone where spends a long time. Then it slowly, over the *depression* zone *Z<0*, returns to regular (oscillatory) mode. This regime qualitatively corresponds to the "*stress*" behavior.

The yellow curve in Fig. 5 represents the *separatrix* between the attraction zone of abnormal stationary state {Z*,μ*} and other behavioral modes. Since this state is also a stable focus, the *affix*, ones getting inside the attraction zone, will be "sucking" up (over several damping oscillation around) to the abnormal stationary state, and cannot leave this zone independently, without serious external impact. It should be stressed that normally, the trajectory cannot cross the separatrix from outside; this could occur only occasionally (due to some small excitation when the affix is near the separatrix). This implies that commonly, the *stress* regime returns to a normal mode and should not result in the shock state. But, since at certain stage of the process the trajectory comes very close to the separatrix, the least excitation could result in hitting the shock zone and fall down to the coma state. Thus, this model enables us to infer that the stress regime is *dangerous* for human beings, since this process includes the stage (just before the stress mode turns to increasing μ values, i.e., to rather normal behavior) when the weak external excitation could provoke momentary stress-to-shock transition. This is *novel* model prediction, which could be tested experimentally. Note that certain evidences in favor of this effect were already detected [59].

This model could be applied to analyze possible results of use of different medical impacts, such as adding certain *stimulants* at different stages of the stress process. Such research could lead to pronounced applied results.

The described effects are in good qualitative agreement with the experimentally observed ones [59]. Quantitative correspondence is intricate, since the characteristics that are measured experimentally are close *per se* to *Z(t)* as a measure of irregularity, and μ(t) as a measure of mediator imbalance. However, there is no direct correspondence between theoretical and experimentally measured variables.

VI. EMOTIONAL MANIFESTATIONS IN COURSE OF INFORMATION PROCESSING

Let us discuss the role of emotions in solving the problems of *recognition* and *prediction***,** which could be accompanied by certain dynamical variation of the noise amplitude *Z(t).* Typical patterns of *Z(t)* behavior will be presented below.

A. Recognition

Note that he extended set of images with distinguished "borders" is needed for good quality of recognition (classification). Usually such classifier is built in course of training the recognition system. Under NCA, **RH** plays the role of Supervisor for **LH,** and that is trained **LH** that implements the function of the object recognition (classification).

The problem of object/phenomenon recognition is solving in already trained system (with at least two trained lower levels σ=0, 1) by means of image plates.

The incoming information is perceived by both subsystems. If this information is well known, the problem is solved in the subsystem **LH** by means of Hopfield-type mechanism of *refinement*: all the images are treated as already known ones — by fitting them to coincide with already stored patterns. In the case of insufficient recognition (when the fitting procedure fails), the participation of **RH** becomes necessary. An unrecognized image is treated as a new one and undergoes the common procedure of new symbol formation.

The problem setting consists in excitation of certain group of neurons ("*examinee* object") in the *fuzzy set* $\boldsymbol{H^0}$ in **RH.** Here, this "object" is processing "as it is", i.e., by *blackening* connections between all the examinee neurons. This neuron group could contain several "skilled" neurons (that belong to certain already known image), with already black connections between them. This means that the examinee object is (to some extent) similar to some familiar (already learned) one.

Then, this image is transferred (by direct cross-subsystem connections Λ) to the typical-image plate $\boldsymbol{H^{typ}}$ in **LH.** Further procedure is controlled by the value of the discrepancy *D*, which could be defined as

$$D(t) \equiv \sum_{i}^{M} \left\| H_i^0 - H_i^{typ} \right\|, \qquad (15)$$

where summation is performed over *M* excited examinee neurons.

There are several possible cases.

1) Familiar object: If the examinee object is well-known to the system, i.e., its image completely coincides with one of typical images in **LH,** so that *D*(0)=0, it would be straight away (quickly!) associated with corresponding symbol, with all the following consequences concerning its position in the hierarchy. In this case, **RH** does not participate further in the process. Accordingly, *dZ(t)/dt=0*,

so this (practically automatic) procedure do nor call for any emotions.

2) Examinee object is close to familiar one: The examinee object can be sufficiently similar to one of the known typical images (fits its "attraction area"), i.e., $D(0) \neq 0 < D_{cr}$, where D_{cr} represents certain critical value of discrepancy. Then, it is treated as familiar one together with its symbol at the next level $G^{L,1}$. However, in this case the recognition propriety requires *verification.* For this purpose, the *symbol* should be transferred to **RH** for decomposition, and the result should be compared with the examinee image. Thus, there arises the *loop*, i.e., iterative process presented in Fig. 6:

$$\begin{array}{ccc} & \overset{\Psi^L}{} & \\ H^{typ}\,(\mathrm{LH}) & \rightarrow & G^{L,1}\,(\mathrm{LH}) \\ \uparrow\Lambda^{R\to L} & & \downarrow\Lambda^{L\to R} \\ H^{0}\,(\mathrm{RH}) & \leftarrow & G^{R,1}\,(\mathrm{RH}) \\ & \underset{\Psi^R}{} & \end{array}$$

Figure 6. Schematic representation of the iterative recognition procedure

If the result of comparison was satisfactory (it could be estimated by running D value), the examinee object would be associated with an existing symbol. If not, the discrepancy provokes repeating, and the procedure should pass over several iterations. At that time, the image in the fuzzy set $\boldsymbol{H^0}$ gradually blackens.

3) Examinee object is far from familiar one: If $D(0) > D_{cr}$, at some moment the connections recording the object in **RH** turn to be sufficiently "black" ($\Omega^0_{ij} \geq \Omega_0$), but the typical-image plate do not recognize the object, — then, it turns out to be the *new typical image* and should take its place at the plate $\boldsymbol{H^{typ}}$, so that $D(t_{D=0}) = 0$. Then, the common procedure of new symbol formation should provide its own symbol, which should be linked to high-level symbols, and so on. The moment $t_{D=0}$ is accompanied by $Z(t)$ decrease — the system had solved the recognition problem and could relax. Typical pattern of $Z(t)$ dynamics in course of recognition procedure is presented in Fig. 7a.

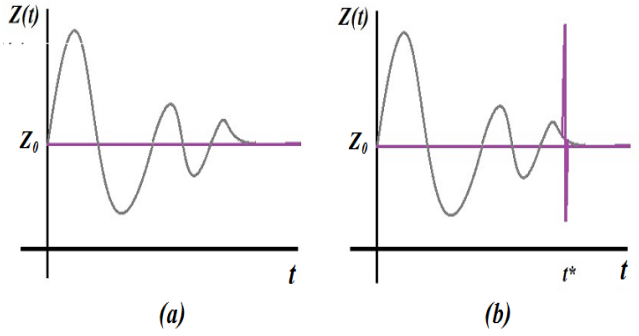

Figure 7. Typical patterns of the noise-amplitude behavior in the cases of (a) recognition procedure; (b) prognosis and incorrect prognosis at the time moment t* (illustration for the sense of humor).

Thus, we can infer that the given system is capable to process and recognize even *new* objects, yet only with participation of the fuzzy set $\boldsymbol{H^0}$.

B. Prognosis

The prognosis (forecast) can be treated as a "recognition of a time-depending process". It proceeds in **LH** after the symbol of the given *process* is formed. This generalized symbol collects all the information about the "process pattern" (image of symbols) in a compressed form. Then, the information on some middle stage of the given process activates it's symbol, providing the activation of the entire chain of symbols enclosed in this process.

Therefore, emotional manifestations, as well as the pattern of noise-amplitude $Z(t)$ behavior here is similar to that in the case of recognition (Fig. 7b). Note that this statement is true up to the moment when the prediction is failed. This means that the information coming at some moment t^* appears to be *unexpected*. This case refers to the problem of the sense of humor (see below).

C. Interpretation of the Sence of Humor

Under the presented concept, the sense of humor could be interpreted as an "*ability to adapt quickly to unexpected information with getting positive emotions*". This process is illustrated in Fig. 8.

Let the incoming data represent a time sequence of symbols that is perceived *consequently* by **LH,** as it is shown in Fig. 8. At the initial stages, the information perceived is usually not concrete enough to correspond to one *symbol of process* at $\boldsymbol{G^2}$, thus the system makes no predictions. A prognosis could be done when accumulated information enables the subsystem to choose one symbol among the others (in Fig.8, "black" symbol at $\boldsymbol{G^2}$ plate, which has more strong connections than the "green" one, i.e., it corresponds to more "common" process). Then, the system *waits* for further details of the predicted process (this means activation of the "black"-symbol chain at $\boldsymbol{G^1}$ plate).

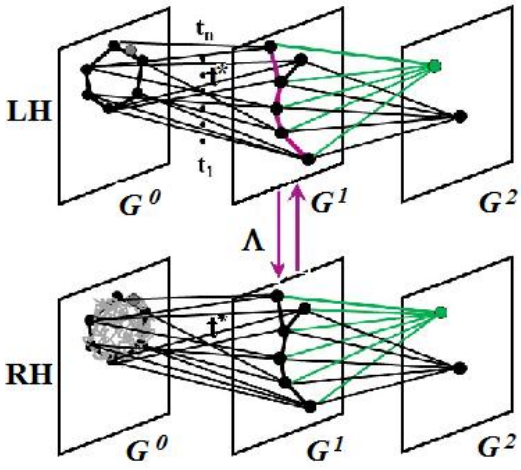

Figure 8. Illustration of the process of perception of incoming information in the well-trained system.

Up to certain moment *t**, the incoming information ("violet" chain in Fig. 8) fits these expectations. At the moment ***t*,*** the prognosis on further information could appear to be *incorrect*, — the next symbol at G^1 plate belonging to "violet" chain, actually is not involved into the "black"-symbol chain, and thus unexpected. Then the system has to appeal to **RH** (down Λ-arrow in Fig. 8); in this process, the emotions are negative: $dZ/dt > 0$. However, the system may rapidly find a new solution — this implies that there *already exists* the symbol of another process that matches completely both, former and current information ("green" symbol at G^2 plate in Fig. 8). This leads to positive emotions ("aha" moment) and hence switches the connections $\Lambda^{R\to L}$ (up arrow in Fig. 8).

According to this concept, a good anecdote should be a story that, up to certain moment t*, permits a well-known interpretation. The next information block should *not deny* the previous version, but suggest another (alternative) also well-known interpretation. In this case, the system has to return to the turning point at ***t**** and then choose the "right" chain of symbols fitting all the incoming information. The very process of returning and jumping to the "right" trajectory requires definite specific efforts — so again it leads to the spike of the noise amplitude that simulates laughter.

Let us stress that all this is possible, if the system is reach enough with symbols of processes, i.e., has large enough "repertoire" of various symbols and images. Then this process is rapid, both trends appear to be superimposed: the value *Z(t)* undergoes abrupt increase-and-decrease ("spike") that could be interpreted as an analogy to human *laughter* (abrupt involuntary reaction). Thus, we infer that a sense of humor could be inherent to the well-trained *(erudite)* system only, just as it is for human beings.

D. Interpretation of Aesthetic Emotions (annons)

NCA could be applied to the problem of analysis of the nature of so called *aesthetic* emotions. Emotions of this type are not connected with any rational (pragmatic) reasons, but are evoked by pure Nature phenomena (rainbow, fire, etc.), pieces of Art, etc. Under NCA, one may suppose that these emotions are associated with the *recognition paradox*: these phenomena seem familiar and surprising simultaneously. In this case, *Z(t)* should display small variations (*vibration*) around the normal level Z_0, that correspond to the human feeling called the "*goosbumps*". In many aspects, the mechanism of aesthetic emotion production is similar to the incorrect\undone prognosis, therefore to the sense of humor. That is why pronounced emotions are often accompanied by the laughter (or tears). However, there is important difference: the unexpectedness in the case of aesthetic emotions could not be "resolved" by switching to another, already known symbol. Nevertheless, this problem deserves further study [60].

VII. PERSPECTIVES ON PRACTICAL VALIDATION

This theoretical study has *per se* fundamental character and could be related to the human-level AI (HLAI) trend. However, its experimental verification represents the most interesting problem.

The comparison of our model predictions with the experimental results on Electro Cardiogram (ECG) analysis under the stress/shock conditions [59] has shown good qualitative agreement. Note that these experiments were based on the analysis of ECG, with *model-dependent* interpretation of the correspondence between ECG pattern and the activity of certain brain areas.

However, the model variable corresponding to the noise amplitude Z(t) has no direct analogues within the experimental technique used. This requires special efforts to extract this information from experimental data on the neocortex activity.

The model predictions concern certain peculiarities in the brain activity, including the cerebral cortex and subcortical structures. The sub-cortical production could be estimated indirectly, by analysis of certain vegetative indices, as it was done, e.g., in [59]. However, meaningful experiments should involve combined study using ECG, Electro Encephalogram (EEG), and functional Magnetic Resonance Imaging (fMRI). In this process, the main attention should be paid to the dynamic variations in the brain-structures activity, thus good enough time resolution of the experimental devices is required.

These techniques are actually available now [61]. We plan to perform such experiments in collaboration with the group of V. L. Ushakov in Kurchatov Research Center, Moscow, in the nearest future. In particular, we plan to perform combined analysis of ECG, EEG, and fMRI data for people under the "light stress" experimental conditions (e.g., time trouble in solving specific cognitive problems).

VIII. CONCLUSIONS AND FUTURE WORK

In summary, the main inference of the paper is that NCA architecture inherently contains the possibility and even necessity to incorporate the emotions into the cognitive process.

The main constructive feature of this architecture is representation of the whole system as a combination of two linked subsystem, **RH** and **LH**, with the presence of random element (noise) in **RH** only. These subsystems could be associated with cerebral hemispheres, with the connections Λ(t) between them representing *corpus callosum.* It was shown that **RH** is responsible for processing the new (therefore, unexpected) information, while **LH** stores and processes the well-known one. This functional specialization is in entire agreement with the practical inferences of Goldberg [19]. The coincidence of theoretical (DTI-based) and practical (practicing psychologist E. Goldberg) inferences represents a pleasant surprise and indirect verification of NCA.

However, this design requires a specific mechanism to control the subsystem activity. It is quite natural to associate this mechanism with the emotional response to incoming information.

It is shown that emotional self-appraisal in an artificial cognitive system could be associated with the variation of the noise amplitude. In order to reproduce human-level

emotional process, which is regulated by the neural transmitters, it should be linked to certain additional variable reflecting aggregated composition of neurotransmitters. Their mutual dynamical interaction in course of cognitive process provides the tool for regulating the activity of subsystems. Thus, the problem of "Explanatory Gap" between the "Brain" and "Mind" approaches to representation of emotions is solved.

Returning to the wide-spread and somewhat "vulgar" idea that **RH** is a "container" for emotions while **LH** provides rational reasoning, we may infer that emotions actually lie *deeper* (in all senses). They belong neither to **RH** nor to **LH**, but actually control their activity. That is why in Fig. 1, "emotions" are virtually displayed beyond both subsystems (associated with *neocortex)*.

The emotional response is described by the *derivative dZ/dt* of the variable that indicates the level of the noise *Z(t)*. Negative emotions imitated by the noise increasing ($dZ/dt>0$) correspond to unexpected incoming information (incorrect and/or undone prognosis, surprise); in this process, **RH** should be activated. Vice-versa, solving any problem results in positive emotions and, correspondingly, decrease of the noise amplitude ($dZ/dt<0$) — then, only **LH** remains active, while **RH** gets an opportunity to be "at rest". Specific case of an abrupt up-and-down jump ("spike") of the function *Z(t)* could be associated with specific human manifestation of emotions (the *laughter*).

Realization of this program in AI could be accompanied by certain sound effects, such as artificial "*laughter"* in the case of abrupt spike of *Z(t)*. In addition, variation of the noise amplitude during the process of problem solving could be accompanied by the display of visual "symbols", such as cheery or sorrowful "faces", etc.

This approach opens a wide field for imitation and model analysis of various human peculiar features. This implies, e.g., that various types of temperament could be associated with certain values of the rest-state noise amplitude Z_0 and thus classified. Furthermore, the model described the stress/shock effect could be employed for working up new medical-treatment techniques for specific (neural) diseases. All these tasks require further study.

It should be stressed that all these possibilities emerge from the human-like cognitive architecture proposed under NCA. Let us accentuate several *key points* of NCA that distinguish it from other neuromorphic approaches and could be applied successfully to artificial cognitive systems (particularly, in Robotics):

- *Continual* representations of neural processors involving nonlinear differential equations.

This representation enables us to interpret and reproduce the experimentally observed effect of mutation of the "*skilled*" neurons (participated in acquisition of certain experience) by the *parametric* modification.

- The whole system represents a combination of *two linked subsystems* (**RH** and **LH**) – for generation and reception of information, respectively.
- *Different training* principles in **RH** and **LH** secure the hemisphere specialization.

New information processing requires the amplification of the new connections (Hebbian principle), while the processing of well-known information (recognition) requires the selection principle "redundant cut-off" (Hopfield's rule).

- Account for *random component* ("noise") presented in **RH** only.

This fact immediately specifies the role of **RH** in the response to unknown/unexpected conditions and leads to:

- Interpretation of emotions as a tool for controlling the subsystem activity, that could be realized via the noise-amplitude derivative *dZ/dt*.
- *Instability* of the image-to-symbol conversion process that leads to unpredictable patterns.

This very factor could secure the *individuality* of an artificial cognitive system.

- The "connection-blackening" principle of self-organization, which provides the possibility for **RH** to acts as a Supervisor for **LH;** no external supervising is needed for permanent learning.

Thus, these design features make it possible to reproduces the peculiarities of *human* cognition — that is, unpredictable character, individuality, permanent learning, ability of logical and intuitive thinking, etc. Note that these problems are actually not considered in other approaches.

It should be stressed that under NCA, the noise (random element) is treated not as unavoidable obstacle (as it is in radio physics, information-delivery tasks, etc.), but as necessary *full member* of all the processes referring to generation of information. Note that the noise (concerning the living systems, this implies fortuitous, spontaneous, sudden act), represents the *survival mechanism* that prevents precise and speed acting (particular for robots) in common situations, but provides an ability to find *occasionally* quite sudden and unpredictable exit from a critical situation. This very factor could provide the human-like features in an artificial system.

Actually, modern AI systems correspond to **LH** under NCA, but this is the **RH** that secures the emergence and individuality of such intellect. Moreover, even in the well-trained cognitive system, the combination of **LH** and **RH** provides rather broad spectrum of abilities than **LH** only — without **RH**, the cognitive system appears to be poor. Thus, we can infer that the NCA architecture, in spite of its seeming complexity and awkwardness, has several advantages comparing with popular AI architectures. Some loss of materials for doubling the system could gain a profit in system's self-development.

It is worth noting that the idea of using two subsystems, with the noise being presented in the one, has already attracted an attention in Robotics [62]. However, this idea requires specific mechanism for switching the activity of certain subsystem depending on the process stage. Under NCA, this mechanism is actually proposed. According to our main hypothesis, it should be controlled by emotions displayed as the noise-amplitude variation.

Thus, it is shown that under NCA, emotional response to external information (including unexpected, i.e., surprising

one) could be imitated by specific behavior of the noise amplitude.

These ideas deserve further research and experimental verification.

REFERENCES

[1] O. D. Chernavskaya and Ya. A. Rozhylo, "On the Possibility to Imitate the Emotions and "Sense of Humor" in an Artificial Cognitive System," Proc. of the Eighth International Conference on Advanced Cognitive Technologies and Applications (COGNITIVE 2016) IARIA, Mar. 2016, pp. 42–47; ISBN: 978-1-61208-462-6.

[2] J. E. Laird, The Soar cognitive architecture, MIT Press, 2012.

[3] A. Samsonovich, "Bringing consciousness to cognitive neuroscience: a computational perspective," Journal of Integrated Design and Process Science, vol. 1, pp. 19–30, 2007.

[4] Y. Bengio and Y. Le Cun, Scaling Learning Algorithms towards AI. In: Large Scale Kernel Mashines, L. Botton, O. Chapelle, D. DeCoste, and J. Weston (Eds.), MIT Press, 2007.

[5] H. Lee, R. Ekanadham, and A. Ng, "Sparse deep belief net model for visual area V2,"Advances in Neural Information Processing Systems (NIPS), vol.7, pp.873–880, 2007.

[6] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, "Deep Learning," Nature, vol. 521, pp. 436–444, 2015.

[7] L. F. Koziol and D. E. Budding. Subcortical Structures and Cognition. Implications for Neurophysiological Assessment. Springer, 2009 .

[8] K. Doya, "Complementary roles of basal ganglia and cerebellum in learning and motor control," Current Opinion in Neurobiology, vol. 10, pp.732–739, 2000.

[9] Springer Handbook of Robotics. Editors: Bruno Siciliano, Oussama Khatib , Springer, 2016. ISBN: 978-3-319-32552-1.

[10] G. A. Carpenter and S. Grossberg, Adaptive resonance theory. In: Encyclopedia of Machine Learning and Data Mining, C. Sammut and G. Webb (Eds.) Berlin: Springer-Verlag, 2016.

[11] O. D. Chernavskaya, D. S. Chernavskii, V. P. Karp, A. P. Nikitin, and D. S. Shchepetov, "An architecture of thinking system within the Dynamical Theory of Information," Biologically Inspired Cognitive Architectures (BICA), vol. 6, pp. 147–158, 2013.

[12] O. D. Chernavskaya, D. S. Chernavskii, V. P. Karp, A. P. Nikitin, D. S. Shchepetov, and Ya.A.Rozhylo, "An architecture of the cognitive system with account for emotional component," Biological Inspired Cogitive Architectures (BICA), vol.12, pp. 144–154, 2015.

[13] H. Haken, Information and Self-Organization: A macro-scopic approach to complex systems. Springer, 2000.

[14] D. S. Chernavskii,"The origin of life and thinking from the viewpoint of modern physics," Physics-Uspekhi, vol. 43, pp. 151–176, 2000.

[15] D. S. Chernavskii, Synergetics and Information. Dynamical Theory of Information. Moscow, URSS, 2004 (in Russian).

[16] J. G. Nicholls, A. R. Martin, and P. A. Fuchs, D. A. Brown, M. E. Diamond, and D. A. Weisblat, From Neuron to Brain, 5th ed. Sunderland, Mass: Sinauer Associates, Inc., 2012.

[17] D. O. Hebb, The organization of behavior. John Wiley & Sons, 1949.

[18] O. E. Svarnik, K. V. Anokhin, and Yu. I. Aleksandrov, "Experience of a First Whisker-Dependent Skill Affects the Induction of c-Fos Expression in Somatosensory Cortex Barrel Field Neurons in Rats on Training the Second Skill," Neuroscience and Behavioral Physiology, July 2015, DOI: 10.1007/s11055-015-0135-3

[19] E. Goldberg, The New Executive Brain. Oxford University Press, 2009.

[20] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," PNAS, vol. 79, p. 2554, 1982.

[21] S. Grossberg, Studies of Mind and Brain. Boston, Riedel, 1982.

[22] T. Kohonen, Self-Organizing Maps. Springer, 2001.

[23] S. Grossberg, "Towards Solving the Hard Problem of Consciousness," Invited talk given at the international conference on Human-Level Artificial Intelligence (HLAI). Available from: http://agi-conf.org/hlai2016/wpcontent/uploads/2016/03/HLAI2016 Retrieved 2016.11.30

[24] J. Levin, "Materialism and Qualia: The Explanatory Gap," Pacific Philosophical Quarterly, vol. 64(4), pp. 354–361, 1983.

[25] J. Stirling and R. Eliott, Introducing Neuropsychology, 2nd ed. (Psychology Focus), Psychology Press, 2010.

[26] R. L. Solso, Cognitive psychology, 6th ed., Pearson, 2000.

[27] Merriam-Webster Dictionary. Available from: http://www.merriam-webster.com/definition Retrieved 2016.11.30

[28] Wikipedia. Available from: http://www.en.wikipedia.org/wiki/Main_Page Retrieved 2016.11.30

[29] E. M. Izhikevich and G.M. Edelman, "Large-scale model of mammalian thalamocortical systems". In: Proceedings of the national academy of sciences (PNAS), vol.105, pp. 9, 2008.

[30] J. Panksepp and L. Biven, The Archaeology of Mind: Neuroevolutionary Origins of Human Emotions. N.Y.: Norton, 2012.

[31] P. Vershure, "The Distributed Adaptive Control: A theory of the mind, brain, body nexus," BICA, vol. 1, pp. 55–72, 2012; "The Distributed Adaptive Control of Consciousness," Lecture given at FIRCES on BICA, April 20–24, Moscow, 2016.

[32] H. Lovehiem, "A new three-dimensional model for emotions and monoamine neurotransmitters," Med. Hypotheses, vol. 78, pp. 341– 348, 2012; DOI: 10.1016./j.mehy.2011.11.016.

[33] J. Valerdu, M. Talanov, S. Distefano, M. Mazzara, A. Tchitchigin, and I. Nurgaliev, "A cognitive architecture for the implementation of emotions in computing systems," Biologically Inspired Cognitive Architectures (BICA), vol. 15, pp.34–40, 2016.

[34] A. Samsonovich "Emotional biologically inspired cognitive architecture," Biologically Inspired Cognitive Architectures (BICA), vol. 6, pp. 109–125, 2013.

[35] M. Sellers, "Toward a comprehensive theory of emotion for biological and artificial agents," Biologically Inspired Cognitive Architectures (BICA), vol. 6, pp. 3–26, 2013.

[36] M. I. Rabinovich and M.K. Muezzinoglu, "Nonlinear dynamics of the brain: emotions and cognition" Physics-Uspekhi, vol. 53, pp. 357–372, 2010.

[37] J. Treur, "An integrative dynamical systems perspective on emotions," Biologically Inspired Cognitive Architectures (BICA), vol. 6, pp. 27–40, 2013.

[38] J. Schmidhuber, "Simple algorithmic theory of subjective beauty, novelty, surprise, interestingness, attention, curiosity, creativity, science, music, jokes," Journal of Science, vol. 48 (1), pp. 21–32, 2009.

[39] D. Huron, Sweet Anticipation: Music and Physiology of Expectation. MIT Press, 2006.

[40] E. Hudlycka, "Affective BICA: Challenges and open questions," Biologically Inspired Cognitive Architectures (BICA), vol. 7, pp. 98–125, 2014.

[41] W. Weaver and C. Shannon, The Mathematical Theory of Communication. Univ. of Illinois Press, 1963; ISBN 0-252-72548-4.

[42] S. Verdu, "Fifty years of Shannon theory". In: Sergio Verdu and Steven McLaughlin: Information Theory: 50 years of discovery, IEEE Press, pp.13–34, 2000.

[43] I. Prigogine, End of Certainty. The Free Press, 1997; ISBN 0684837056.

[44] H. Quastler, The Emergence of Biological Organization. New Haven: Yale University Press, 1964.

[45] T. W. Deacon, The Symbolic Species: The Co-Evolution of Language and the Brain. Norton &Co, NY, 1997.

[46] E. M. Izhikevich, Dynamical systems in neuroscience: the geometry of excitability and bursting. MIT Press, 2007.

[47] A. L. Hodgkin and A. F. Huxley, "A quantitative description of membrane current and its application to conduction and excitation in nerve," Journal of Physiology, vol. 117, pp. 500–544, 1963.

[48] R. FitzHugh, "Impulses and physiological states in theoretical models of nerve membrane," Biophys. J., vol. 1, p. 445, 1961.

[49] J. Nagumo, S. Arimoto, and S. Yashizawa, "An active pulse transmission line simulating nerve axon," Proc. IRE, vol. 50, p. 2062, 1962.

[50] V. L. Bianki, "Parallel and sequential information processing in animals as a function of different hemispheres," Neuroscience and Behavioral Physiology, vol. 14 (6), pp. 497–501, 1984.

[51] C. M. Bishop, Pattern Recognition and Machine Learning. Springer, 2007.

[52] W. S. McCulloch and W. Pitts, "A Logical Calculus of the Ideas Immanent in Nervous Activity," Bulletin of Mathematical Biophysics, vol. 5, p. 115, 1943.

[53] J. R. Anderson, How can the human mind occur in the physical universe? Oxford University Press, NY, 2007; ISBN 0-19-532425-0.

[54] C. L. Dancy "ACT-R<PHI>: A cognitive architecture with physiology and affect," Biologically Inspired Cognitive Architectures (BICA), vol. 6, pp. 40–45, 2013.

[55] The Cerabral Cortex Map. Available from: http://www.Antranik.org/functional-areas-of-the-cerebreal-cortex / Retrieved 2016.11.30

[56] I. Kant, Critick of Pure Reason, ed. William Pickering, London, 1838; original 3-d edition: Kritic der reinen Vernunft, ed. J.F. Hartknock, 1790 (in Deutch).

[57] O. D. Chernavskaya, D. S. Chernavskii, and A. P. Nikitin, "The concept of intuitive and logical in neurocomputing," Biophysics, vol. 54, pp. 727–735, 2009.

[58] G. A. Wiggins, "Learning and Creativity in the Global Workspace," Advances in Intelligent Systems and Computing, vol. 196, p. 57, 2013.

[59] S. B. Parin, A. V. Tsverlov, and V. G. Yakhno, "Models of Neurochemistry Mechanism of Stress and Shock Based on Neuron-like Network," Proceedings of Int. Simp. "Topical Problems of Biotonics," Aug. 2007, pp. 245–246.

[60] O. D. Chernavskaya, D. S. Chernavskii, Ya. A. Rozhylo, and D. S. Shchepetov, "Hypothesis on the nature of aesthetic emotions and the concept of Chef-d'eouvre" (unpublished).

[61] A. S. Sedov, D. A. Devetiarov, U. N. Semenova, V. V. Zavyalova, V. L. Ushakov, R. S. Medvednik, M. V. Ublinsky, T. A. Akhadov, and N. A. Semenova, "Dynamics of Brain Activity during Voluntary Movement: fMRIStudy," Zh. Vyssh. Nerv., vol. 65, pp. 436–445, 2015;
DOI:10.7868/S0044467715040115

[62] K. Kushiro, Y. Harada, and J. Takeno, "Robot uses emotions to detect and learn the unknown," Biologically Inspired Cognitive Architectures (BICA), vol. 4, pp. 69–78, 2013.