Psychology in Russia: State of the Art
Volume 10, Issue 3, 2017

Russian Psychological Society
Lomonosov Moscow State University

A computational model of consciousness for artificial emotional agents

Artemy A. Kotov

Kurchatov Institute National Research Center, Moscow, Russia
Russian State University for the Humanities, Moscow, Russia

Corresponding author. E-mail: kotov_aa@nrcki.ru

Background. The structure of consciousness has long been a cornerstone problem in the cognitive sciences. Recently it took on applied significance in the design of computer agents and mobile robots. This problem can thus be examined from perspectives of philosophy, neuropsychology, and computer modeling.

Objective. In the present paper, we address the problem of the computational model of consciousness by designing computer agents aimed at simulating "speech understanding" and irony. Further, we look for a "minimal architecture" that is able to mimic the effects of consciousness in computing systems.

Method. For the base architecture, we used a software agent, which was programmed to operate with scripts (productions or inferences), to process incoming texts (or events) by extracting their semantic representations, and to select relevant reactions.

Results. It is shown that the agent can simulate speech irony by replacing a direct aggressive behavior with a positive sarcastic utterance. This is achieved by balancing between several scripts available to the agent. We suggest that the extension of this scheme may serve as a minimal architecture of consciousness, wherein the agent distinguishes own representations and potential cognitive representations of other agents. Within this architecture, there are two stages of processing. First, the agent activates several scripts by placing their if-statements or actions (inferences) within a processing scope. Second, the agent differentiates the scripts depending on their activation by another script. This multilevel scheme allows the agent to simulate imaginary situations, one's own imaginary actions, and imaginary actions of other agents, i.e. the agent demonstrates features considered essential for conscious agents in the philosophy of mind and cognitive psychology.

Conclusion. Our computer systems for understanding speech and simulation of irony can serve as a basis for further modeling of the effects of consciousness.

Keywords: cognitive architectures, psychophysiological problem, theory of consciousness, emotional computer agents, machine humor, simulation of irony, text comprehension

ISSN 2074-6857 (Print) / ISSN 2307-2202 (Online)
© Lomonosov Moscow State University, 2017
© Russian Psychological Society, 2017
doi: 10.11621/pir.2017.0304
http://psychologyinrussia.com

Introduction

For many authors, the notion of "consciousness" is not a strong scientific concept, but rather an element of a "naïve world picture" (e.g., Bulygina & Shmelev, 1997). Usually consciousness is described as a subjective space, which holds mental processes (percepts, representations, thoughts), and is available for observation in the same way as the surrounding physical world. Often a person refers to this space as their "self" ("me" or "I"), although this notion can also refer to an internal world in a wider sense: to one's own knowledge, principles, and, of course — one's own body. "Self" (or consciousness) is also considered to be a source of voluntary actions. When a person acts automatically or reflexively, it is usually suggested that not only the stimulus but the reaction itself resides beyond the boundaries of consciousness, as if something external imposes the reaction on people. However, if a person acts "rationally" or "deliberatively", it is considered that consciousness is the source of these acts.

The "naïve view" of a person with regard to his or her own emotions is more complicated. On the one hand, people usually attribute their own emotional actions to rational choice. Metaphorically, emotions are seen to constitute an external force which leads a person to execute a certain action. On the other hand, one usually highly evaluates one's own emotions, according them priority with regard to any ensuing choices and the evaluation thereof. One can even say: *I understand, but intuitively* ***I*** *feel different* or *I know what I ought to do, but* ***I*** *want to do something different*. In these cases one addresses one's own emotions as a the "true Self". At the same time, the machinery of emotions is not consistent. M. Minsky (1988, p. 165) introduced the concept of "proto-specialist" — a simple model of emotions and drives for an "artificial animal". Each proto-specialist is responsible for the detection of a dangerous (or lucrative) situation and competes with other proto-specialists in order to force the body (the whole organism) to execute a suggested action. The balance between proto-specialists (or other mental agents) will constitute the central point for our further study of consciousness.

As the notion of consciousness is subjectively evident but hard to address scientifically, numerous approaches to this problem have emerged (see, e.g., Chernigovskaya, 2016; Velichkovsky, 2015). In the philosophy of mind, the notion is linked to studies of *understanding* — consciousness is frequently considered as an "organ" for understanding: a mental processor or container for the understood meaning. J. Searle (1980), in his "Chinese room argument", examined and criticized a theoretical design of an *understanding* computer agent — a digital computer. Searle argued that a digital computer solely operates with the data, following the defined rules, and implied that the whole model (and any computer) could not achieve the skill of understanding. In a similar way, T. Nagel (1974) argued that consciousness is incognizable, as no technology (imitation or physical transformation) can let us know what it is like to be a living being — *a bat*. Modern approaches shift the emphasis in this classic discussion: computer models of understanding (speech processors, robot behavior planners) are designed to operate with texts or with behavioral patterns. They are not intended to "make us feel like a robot" nor to "show us the modeled consciousness". So the model of consciousness cannot be falsified, if it does not immerse us in the modeled consciousness, just as engineering models

are not designed to "make us feel like a bridge", but rather to test the bridge in different situations.

Another major approach to consciousness is the attempt to describe introspection or self-awareness. It is suggested that introspection is either essential for consciousness or is a form of consciousness and thus the simulation of introspection may give us a clue to the simulation of consciousness. An analysis of these theories was recently conducted by M. Overgaard and J. Mogensen (2017). A theoretical model of introspection usually has a "double-layer" architecture, where the first layer is responsible for general cognitive tasks, and the second layer monitors or alters the first layer. In a procedural approach undertaken by A. Valitutti and G. Trautteur (2017), it is suggested that on the first level, a system runs general cognitive tasks, while the second level may inspect and alter these basic operations. An example is a software interpreter, which executes the code, simulates the execution (traces and mirrors the code), and may insert additional instructions based on the examination of a single instruction (*local introspection*), as well as on the entire target program (*global procedural introspection*) (Valitutti & Trautteur, 2017). Half a century ago, M. Minsky (1968) proposed that a living being (a man — M) may have a model of self, M*, which answers questions like "how tall am I?" — and a higher level model, M**, with descriptive statements about M*. Minsky suggested that the distinction between M* and M** leads to a "body and mind" paradox, whereby one cannot explain the interaction between cognition and the brain — as mental and physical structures are natively represented by different models.

Although the "double-layer" architecture is widely used in theoretical studies and computer simulations, the definition of introspection as an essential attribute of consciousness may limit the model: subjectively we may be "conscious" when acting in the real world and thinking about real objects — not only at a time of introspection. Therefore, the model should be elaborated to suggest the state of consciousness in different situations, not only in the state of self-awareness.

In psychology, consciousness is frequently explained by the notion of *short-term memory*. It is suggested that short-term memory is the machinery supporting the mental structures which we subjectively perceive to be the content of consciousness. The computer metaphor, applied here to living creatures, indicates the amount of information (objects, features etc.) that can be simultaneously preserved and processed by the subject (for details of this concept, see B.B. Velichkovsky, 2017). It might be that "simple" creatures have a limited *memory*, reducing their behavior to simple reactions. On the other hand, humans have an extended *memory,* allowing them to operate with language structures, mental images, logical inferences, etc. Adherence to this latter metaphor brings us to some questionable results. Modern computers have a huge amount of RAM accessible by software. This however does not evolve them to a threshold of gradual emergence of consciousness, suggesting that the mode of operation may be far more significant than the amount of data processed.

Theoretical approach

Following the analysis of the "naïve" notion of consciousness, we may define a list of features to be modeled by software to produce a "conscious" agent (if a mod-

el of consciousness is indeed possible). A computational model of consciousness should:

- provide space for subjective imaging[1] including pretend images, and establish some kind of coordination between images for further goals;
- generate and verify subjective images or intentions;
- distinguish "self" and "non-self" images, inferences, feelings, or intentions, classify subjective images, and attach subjective feelings to the images;
- handle and possibly solve clashes between conflicting images, feelings, or intentions;
- generate and coordinate actions in a sophisticated way — not only on the basis of pure reactions, but with the consideration of many significant factors

We rely on a cognitive architecture, developed within the Cognition and Affect project (CogAff). This is a "shallow" cognitive model, designed to depict basic cognitive and emotional functions and to be implemented by virtual computer agents (Sloman, 2001; Sloman & Chrisley, 2003). CogAff architecture relies on a "triple-tower" model by Nilsson with a perception module receiving data from the environment, a central processor, and an action unit responsible for the generation of actions (Nilsson, 1988). On different levels, CogAff distinguishes: (a) procedures for emotional processing — *alarms* or *reactions*, (b) deliberative reasoning — models for rational inferences, and (c) models for reflective processes on a "meta-management" level. Entities on each level compete in processing information and in generating output. In CogAff architecture, a lower level of emotional reactions is separated from rational processing by an attention filter. Processes under the attention filter are executed automatically. They can stay removed from attention and consciousness, only to inform the deliberative processing level that a certain reaction took place. At the same time, cognitive structures above the attention filter belong to the deliberative reasoning (or meta-management) levels and simulate the reasoning process of human consciousness. CogAff agents effectively handle some important tasks, like solving conflicts between emotional and rational processes. The architecture also suggests the concept of "tertiary emotions", which use meta-management to inject mental images that have originally driven an emotional response — as in the case of phobias and longing — so that the agent frequently returns to the emotional stimulus in its "thoughts".

With all the advantages of the model, developers can rely on the labels attached to model levels to define "deliberative" processes or "consciousness". Unfortunately, simple labeling of different levels does not explain the structure of consciousness: if a process operates on the level labeled as "consciousness", this does not imply that the process is innately *conscious*. Instead, we have to suggest a specific architecture, operating with different mental objects and sufficiently elaborate to represent an "architecture of consciousness". On the way to the definition of this architecture,

[1] "Images" are understood in the present context as visual, auditory, spatial, linguistic, and emotional representations.

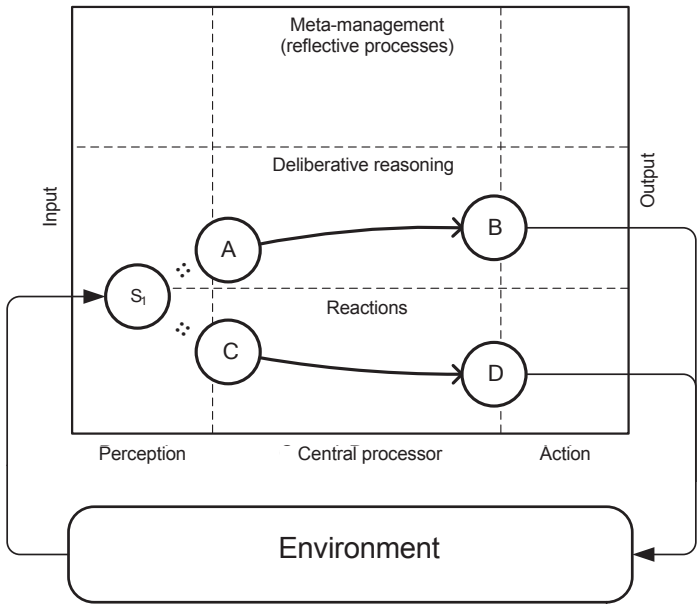


Figure 1. CogAff (Cognition and Affect) scheme as a shallow model of a software cognitive agent

we may suggest several alternatives on how the natural consciousness might be designed. There are the following possible options:

(a) Human consciousness is located in some "spiritual" world and is not connected to any physical (biological) substrate of the body. In this case, all scientific studies of the brain are useless because consciousness cannot be implemented in any hardware or software architecture.

(b) Human consciousness resides in some elements of the brain — molecules, proteins, or other units — and is explained by their physical features. In this case, consciousness cannot be implemented on any hardware, but only on the natural brain tissues or neural network.

(c) Human consciousness is a structural scheme, a mechanism for the interaction of ideal or physical entities. In this case there might be a possibility to implement consciousness with the help of a computer model, relying on existing or future algorithms.

In our view, option (a) does not meet the law of parsimony — even if consciousness has an ideal nature, this option can be preferred only if all conceivable approaches within (b) and (c) are exhausted. Option (b) has an immediate relation to the psychophysiological problem, and suggests that consciousness stems from specific physical (chemical or biological) elements within the brain. If these elements form some structural schemes, suggesting a machinery of consciousness, then these schemes can be modeled by theoretical or real computer architectures — and we arrive at option (c). However, if consciousness is connected to some immanent

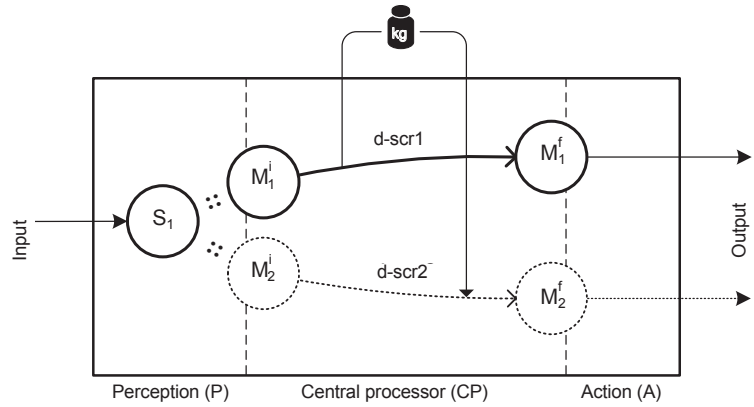
features of the physical brain (as the feature "golden" is connected to the nature of the mineral "gold"), then we arrive at the paradoxical inference that consciousness is a characteristic of matter. Following these inferences, we choose option (c) as the most substantiated. This option suggests that consciousness is a structural scheme, implemented in the physical machinery of the brain. It can be generally described via a theoretical model and run on data processors with various hardware. This approach also suggests that consciousness (or *the effects of consciousness*) can be studied and modeled even before the "psychophysiological problem" is solved.

If we follow option (c), we should roughly assess the number of elements within this architecture. It is usually expected that computer models of consciousness should simulate physical brain structure, and thus should operate with with the scale of the whole brain and not by a structure with fewer elements. We shall follow the opposite approach, however, and suggest that there does exist a *minimal architecture of consciousness*, which is simpler than that of the entire human brain. In the present publication we shall represent our view of the key features of this minimal architecture. We rely on a theoretical model operating with scripts and its computer implementation suggesting that consciousness or *the effects of consciousness* appear if an agent has the capacity to process one stimulus simultaneously with a number of scripts, and if a subsequent script during its activation can access a set of scripts at the previous level. A key example of our approach is the computer simulation of irony.

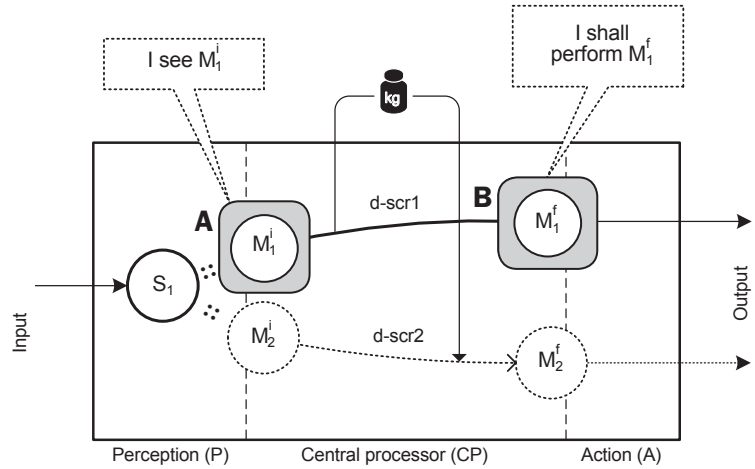
The model of consciousness

In many approaches it is suggested that an emotional analysis of input competes with rational (conscious) processing. So the procedure of emotional text processing may be a key to the understanding of the architecture of consciousness. Earlier, we (Kotov, 2003) presented a list of *dominant scripts* (*d-scripts*), responsible for the recognition of emotional patterns in a natural text, and competing with rational procedures (*r-scripts*) during input processing. A script is a sort of production (inference) with an if-statement — *initial model* and action — *final model.* The list of negative d-scripts consists of 13 units responsible for the recognition of patterns: *It affects your health; They will kill you* (DANGER d-script); *There is no way to go* (LIMIT); *They are just crazy* (INADEQ); *Nobody needs you* (UNNEED); *Everything is useless* (FRUSTR), etc. These scripts appear in dialogues involving conflict (*You don't even care if I die!*) and in negative propaganda (*The government does not care!*). The list of positive scripts includes 21 units for: *It is beautiful* (VIEW); *This sofa is so nice and cozy* (COMFORT); *You control the situation perfectly* (CONTROL); *Everybody loves you* (ATTENTION), etc. These scripts appear in compliments, advertising, and positive propaganda.

A computer agent operated by *d/r-scripts* proved to be capable of simulating speech irony (Kotov, 2009). The agent acted in the following way: when receiving input about an event such as "Someone is hitting you", it activated a negative script DANGER and was ready to reply, *I was hit! You — idiot!* However the agent was suppressing the direct expression of DANGER script in speech; instead, it was looking for a positive script with the highest level of activation — this was the ATTENTION script, usually expressed in the utterances *It's a good thing you have paid*



a



b

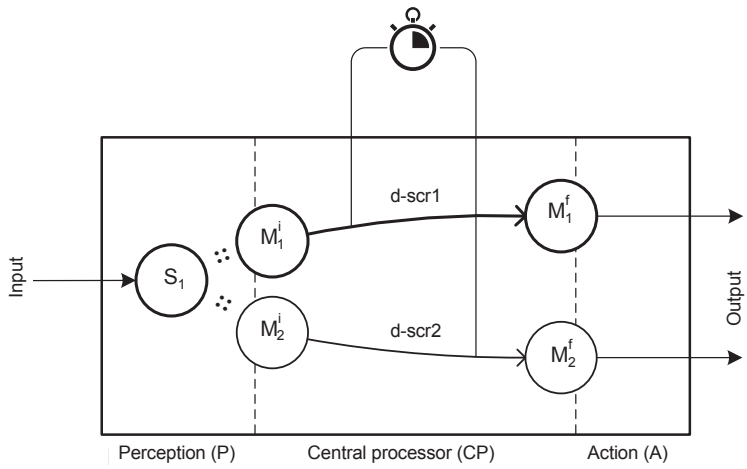
Figure 2. Architecture of an agent inhibiting alternative reactions

attention to me! It's a good thing you care about me! The agent used the utterances from ATTENTION to express the concealed activation of the DANGER script, adding to the utterances a marker of irony (see further details in Figure 5).

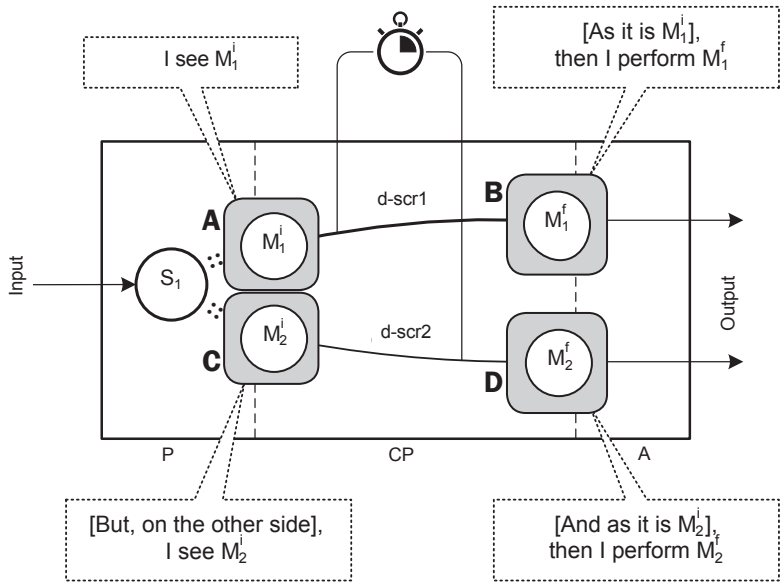
A balance between different scripts forms the cornerstone of our approach to the *minimal architecture of consciousness*. Let us see how this balance is achieved during processing of a stimulus in simple reactive architectures, having no relation to conscious processing (Figure 2). A stimulus S_1 may activate a number of scripts — in particular d-scr1 with high activation, and d-scr2 with lower activation. In Figure 2a we demonstrate an architecture that selects the winning script through script displacement (inhibition of scripts with lower activation). If d-scr1 has received higher activation, then an alternative d-scr2 is suppressed and never appears in the output (indicated by a dotted line).

Quite frequently the notion of consciousness is explained through the notions of *operative memory* and *attention*. We shall use the term *scope of processing* (or

processing scope) in a similar sense. We affirm that the processing scope may contain initial and final models of the scripts. We can compare the processing scope to a desktop with work materials: in order to add any new material, we have to clear space on the desk and remove some older papers. Any inference can be made only on the basis of materials already on the desktop. All papers once removed from the desk no longer exist and are not accessible for immediate cognitive operations. We note that the processing scope of a simple agent (as in Figure 2) contains only one script model. Then, for the agent in Figure 2, the processing scope initially will be



a



b

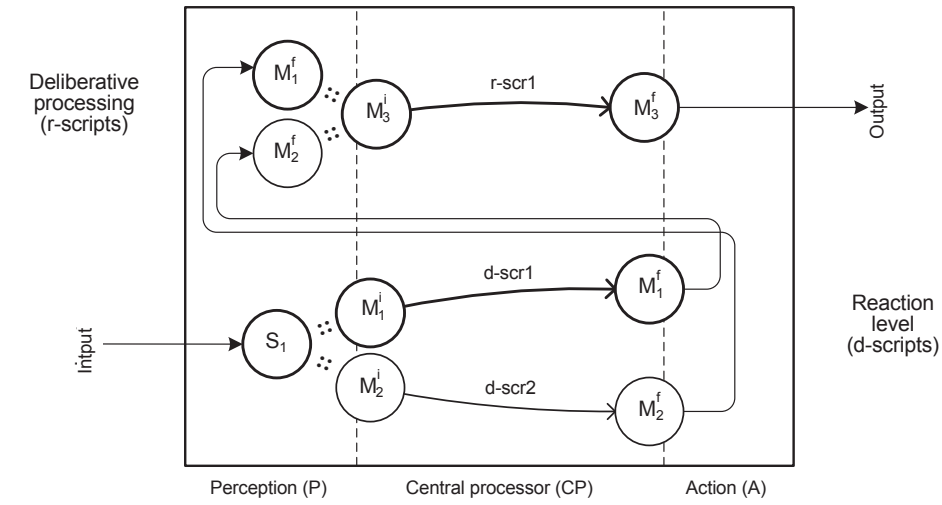
Figure 3. Architecture of the agent with temporal distribution of scripts

in position A, and contain M_1^i model (Figure 2b) — at this stage the agent interprets a stimulus S_1 as M_1^i and believes that M_1^i takes place in the reality in front of him. While activating script d-scr1 and moving to the final model of the script, the agent replaces the contents of the processing scope: in B position the whole processing scope is occupied by the M_1^f model. If, for example, the initial model M_1^i had the content "Somebody is hitting me", then the final model M_1^f may provoke the responsive aggression of the agent. An alternative script d-scr2 will then be inhibited and will never be used to react to the S_1 situation.

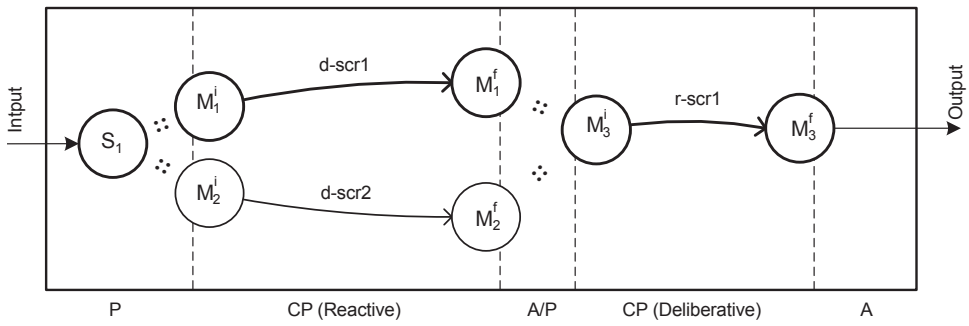
The reactions of an agent may be distributed in time (Figure 3). In this situation, the input S_1 will activate scripts d-scr1 and d-scr2 (as in the previous case). First, d-scr1 will take place, as a script with higher activation, while d-scr2 will be temporarily suppressed. Second, d-scr2 will take place after a standby period. For example, if we "step on the foot" of the agent, he may, first, curse, and second, suggest a socially acceptable reply, saying, *It's all right!* We used the temporal distribution of scripts to simulate the spoken emotional behavior of a computer agent (A. Kotov, 2007). In this architecture, the processing scope will sequentially reside in the A, B, C, and D positions — Figure 3b. First the agent interprets S_1 as $A(M_1^i)$. The interpretation M_2^i at this moment is also constructed by the agent, but this is temporally delayed and no longer remains within the processing scope. Then, the agent reacts to M_1^i — moving to B position and executing actions as defined by M_1^f. When the d-scr1 script is completely processed, the agent shifts to d-scr2. Now it moves representation M_2^i to the processing scope (position C) and then proceeds along with d-scr2 to the inferences or actions of M_2^f (position D). The agent may lack the resources to discover the co-reference of $A(M_1^i)$ и $C(M_2^i)$ so as to understand that these are two different representations of the same situation S_1. If the processing scope presents only one model, then M_1^i и M_2^i will never appear at the processing scope simultaneously so as to be compared by the agent — and the agent will not discover their partial similarity and co-references. As for the result of this limitation, the agent may construct contradictory representations of one and the same situation — and react accordingly to these representations.

Agents shown in in Figures 2 and 3 have very simple architectures: they use scripts from only one level of processing (d-scripts) and can place at the processing scope only one model. More sophisticated agents combine the reactive level with deliberative processing and can activate both d-scripts and r-scripts. In CogAff architecture, this situation can be represented as seen in Figure 4a: final models M_1^f and M_2^f of d-scripts d-scr1 and d-scr2 from the action component of the reaction level are transferred to the input of the deliberative processing level, and may activate a rational script — r-scr1.

We shall rearrange this scheme and draw the processing cycle as a straight line (Figure 4b). Let the scripts d-scr1 and d-scr2 reside on the left from r-scr1, while at the same time keeping in mind that they belong to two different levels of processing: reactive and deliberative. R-script r-scr1 can be activated by M_1^i (then M_1^f is interpreted as M_3^i) or by M_2^f (then M_2^f is interpreted as M_3^i). If the r-script is activated by one of these models, then we can get similar architectures with inhibition or temporal distribution of scripts — as we have seen before (Figures 2 and 3). The main difference is that these architectures work on the upper — deliberative — level. However, we have to pay attention not to the sequential processing, but to



a



b

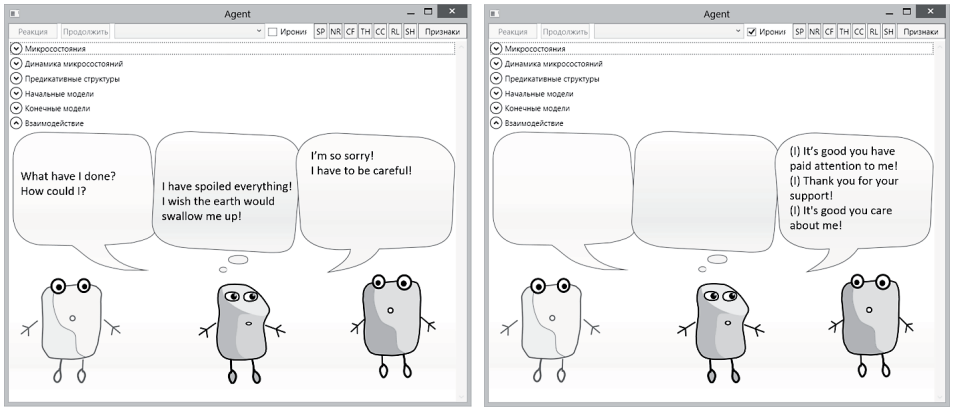
Figure 4. Two-level architectures, distinguishing reactive and deliberative processing levels

the simultaneous processing of competing scripts. Consider architectures where scripts and procedures on upper levels have access to several scripts activated on a previous level. In particular, such a mechanism provides a machinery for irony and ironic replies. Irony for us constitutes a significant example, as it is usually considered to be a sophisticated cognitive task, requiring strong conscious processing.

Computer model of irony

Earlier, we represented a computer agent simulating irony with the help of d/r-scripts (Kotov, 2009). In Figure 5a we show an interface where a Green computer agent (at the center) interacts with other agents: Yellow (on the left) and Grey (on the right). Green receives different predicative structures at its input — these can be system events generated by certain system states, interaction with a user (e.g., by mouse clicks) or semantic components constructed by a syntactic parser as a result of natural text analysis. In a case of ironic behavior (Figure 5b), the agent receives an event "Green (other) is hitting Green (self)", evaluates this event as negative, but

suppresses output of curses, and replies ironically: *Thank you for your support!* and *It's a good thing you care about me!* The ironical nature of the text is indicated by the (I) marker in the interface.



a **b**

Figure 5. Emotional computer agents, software interface

The software processor of the computer agent contains a number of scripts — positive and negative d-scripts responsible for emotional reactions, and r-scripts responsible for rational and socially acceptable replies. The agent compares each incoming event (semantic predication) with the initial models of scripts, calculates the degree of similarity, and defines the activation level for each script. Then the scripts are sorted by the degree of activation. The most activated scripts obtain control over the agent: the agent will then perform gestures and output utterances, as defined for that script.

Following Table 1, d-script DANGER gets the highest activation, 4.1097, after processing the "Somebody hit me" event. If this script gets control over the (Green) agent, the agent complains and swears that "He has been beaten", or shouts at the counterpart Grey agent. Instilling irony, the agent suppresses direct expression of the winning negative script and chooses a positive script with the highest activation. As seen in Table 1, these scripts are CARE (5th line), ATTENTION (6, 12, and 13th lines) and COMFORT (15th line). They all are accorded a similar degree of activation, 2.3482, almost twice as low as that of DANGER (4.1097). From the point of view of the agent, — DANGER is the most relevant classifier (script) for the situation "Somebody hit me"; however, the agent has the ability to choose a positive script with the highest activation to output an ironic answer. ATTENTION type 1 with output utterances *Thank you for your support!* and *It's a good thing you care about me!* was among others selected by the agent in our first experiments. Following the activation level, ATTENTION is not a relevant class (script) for the initial stimulus and can be used only as an extension or a substitute to express DANGER. The initial model of the ATTENTION script is not "what actually takes place" (because some "danger" takes place) and not "what the agent actually feels" (because the agent feels the "danger" — "fear" or "aggression"). Yet this classification of the initial stimulus is still preserved and may be used in a

Table 1. Activation of scripts for an event "Somebody is hitting me"

No.	Score	Script	Possible output
1	4.1097	DANGER	*You will kill me!*
2	3.3482	LIMIT	*You limit me!*
3	3.3482	SUBJECT	*You like to command!*
4	2.3482	PLAN	*You meant that!*
5	2.3482	CARE	*You care about me!*
6	2.3482	ATTENTION type 1	*It's good you have paid attention to me!* *Thank you for your support!* *It's a good thing you care about me!*
7	2.3482	RULES type 1	*It is all right!*
8	2.3482	RULES type 2	*What shall I do in return?*
9	2.3482	Reconciliation	*It is for the best!*
10	2.3482	INADEQ type 4	*You are an idiot!*
11	2.3482	INADEQ type 5	*You don't know what you are doing!*
12	2.3482	ATTENTION type 2	*You are great!*
13	2.3482	ATTENTION type 3	*You understand me!*
14	2.3482	DECEIT	*You lie to me!*
15	2.3482	COMFORT	*I feel great!*
16	2.0133	EMOT	*You are hysterical!*
17	2.0133	SUBJECT type 1	*You think only about yourself!*

communication. This is possibly because the ATTENTION script was not inhibited by other scripts, and the mechanism of irony could access this script among other possible reactions. It means that the processing scope should have some minimal size (here — 15 scripts), to contain a list of scripts with similar or higher activation compared to those suited for the activation of ATTENTION type 1; this ensures that the mechanism of irony can select a suitable "ironic" reaction from the processing scope. If the processing scope contains a number of scripts with different levels of activation, then the agent may differentiate the scripts as "more/less relevant to the situation" or as "my own reactions"/"possible reactions" — this choice can be made simply by means of the activation level. In previous architectures (Figures 2, 3), script activation itself indicated the relevance of the script. Alternatives with lower activation (less relevant scripts) were inhibited or delayed. The agent did not have to compare scripts depending on their activation — this function was effectively executed by an inhibiting process or by a timer. However in the case of irony represented here, the processing scope maintains the ATTENTION type 1 script, which has quite a low activation (not a relevant factor), is neither inhibited nor delayed, and can be accessed by a special communication strategy — making use of irony.

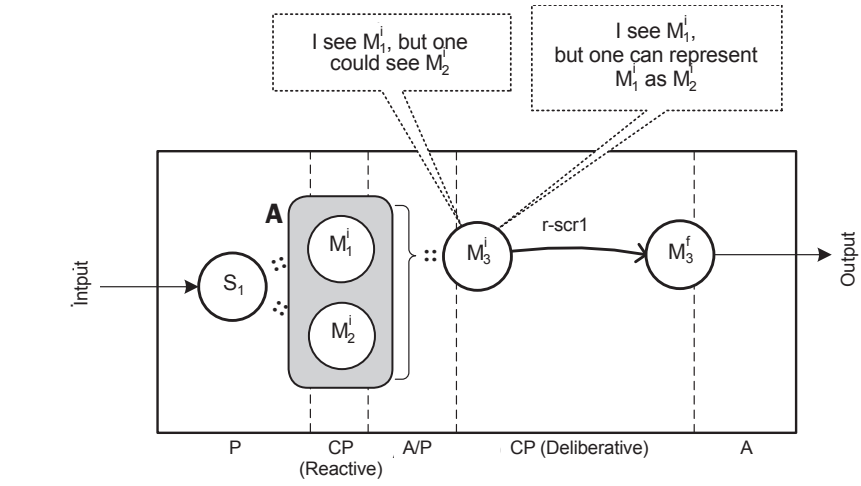
An important feature of the irony mechanism is that it distinguishes (a) a "true" script, which corresponds to the situation and the agent's feelings (in particular, DANGER in the situation of aggression), and (b) an "ironic" script, targeted at the addressee and not reflecting the agent's "true" feelings. Thus the procedure of irony obtains access to two scripts of different degrees reflecting the inner world (or "self") of the speaker and opposite in their evaluation of the situation. Thanks to this architecture, higher-level scripts (or other processing mechanisms, such as irony) can observe the conflict between scripts activated on a lower level and select the scripts that best correspond to the *self* of the speaker. In our example DANGER will better correspond to the speaker's *self* — if we understand *self* as a subjective emotional evaluation, and the ATTENTION script will be targeted at the communication in order to conceal real emotions or to obfuscate the social (communicative) image of the speaker.

In one of the versions of our computer agent we have limited the list of processed scripts to 4 in order to reduce memory load. This change switched off the ability of the agent to synthesize ironic utterances. A negative event could activate several negative scripts, which occupied all 4 slots in the processing scope. In this case the "best" positive script was left out of the allocated memory and could not be accessed through the mechanism of irony. The extension of the processing scope allows the agent to choose the most relevant positive script in a negative situation (and vice versa), even when "top memory slots" are occupied by negative d-scripts, more relevant in a negative situation.

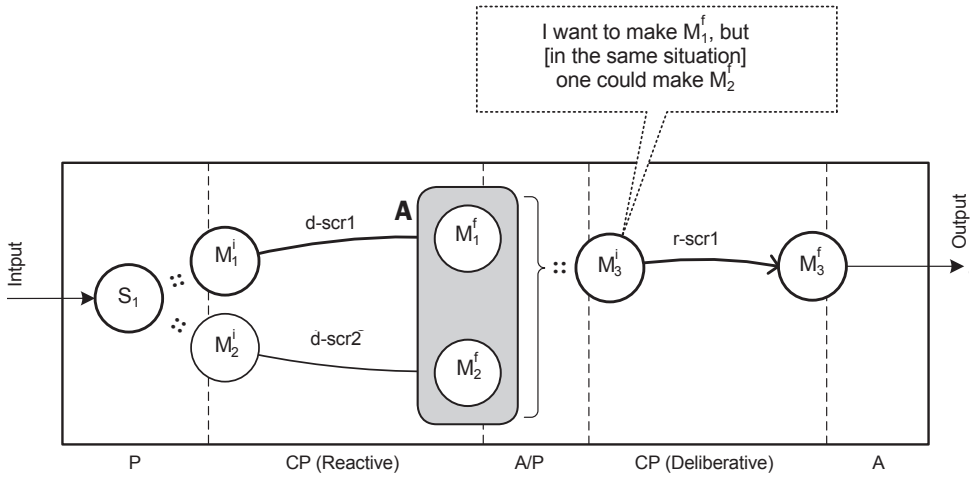
Implication for the model of consciousness

In general, the architecture of irony, and possibly consciousness, requires that: (a) a set of scripts is maintained simultaneously in the processing scope and (b) a further r-script (or mechanism of irony) "sees" these scripts — thus gaining access to many scripts in the processing scope — and is able to distinguish these scripts depending on their activation. This architecture is represented in Figure 6.

Let an incoming stimulus S_1 activate scripts d-scr1 and d-scr2, where both initial models of these scripts are kept in processing scope A. Let us consider the situation whereby an r-script gains access directly to the initial models of these scripts — Figure 6a (final models of this scripts are not shown on the figure). In the processing scope, model M_1^i obtained higher activation, and model M_2^i lower activation. If script r-scr1 can detect this distinction, then the agent "knows" that a situation M_1^i is taking place (the agent "sees" M_1^i); however, in this situation one could see M_2^i. For the agent, it means that it has both "real", as he believes, representation M_1^i and an alternative representation M_2^i (or even a set of such representations). In particular, M_2^i may be used for irony, for the representation "in another situation I could see here M_2^i", "this situation can be represented as M_2^i", "somebody else can see here M_2^i". Thus the extension of processing scope and the ability of r-scr1 to distinguish models in this scope allow the agent to construct a range of "more real" and "more fantastic" representations of an initial S_1 stimulus. The agent thus becomes capable of distinction between reality and alternative representations of reality.



a



b

Figure 6. Architecture of the computer agents for simulation of irony and *the effects of consciousness*

Now consider another situation, where an r-script gets access to the final models of d-scripts — Figure 6b. Here models M_1^f and M_2^f are inferences from an initial situation, actions to be executed by the agent, or goals to be achieved. In any case, M_1^f and M_2^f reflect possible reactions of the agent to the initial stimulus S_1. As in the previous case, the agent may start to distinguish M_1^f and M_2^f as alternative reactions to the situation S_1 if the following conditions are satisfied: (a) processing scope A is big enough to contain M_1^f and M_2^f; (b) r-scr1 has access to both M_1^f models; and the agent can identify the models as alternative reactions to S_1 and at the same time can distinguish the models, based on some differential semantic features. M_1^f obtains higher activation than M_2^f, as d-scr1 initially was more highly activated than

d-scr2. While observing the difference in activation level, the agent may conclude that M_1^f is the main reaction to S_1, and M_2^f is an alternative reaction, suitable in the following situations:

- "In a bit different situation *I* could decide/make M_2^f";
- "In a bit different mood/state *I* could decide/make M_2^f";
- "Somebody else in this situation could decide/make M_2^f".

Thus, observing M_1^f and M_2^f, the agent may conclude that some of the available reactions correspond to its *self* (M_1^f), while other reactions are alternatives that less precisely correspond to its *self* (M_2^f) — they can apply to different situations or to different subjects. In other words, in the range M_1^f, M_2^f, …, M_n^f the agent observes the difference between *self* and *non-self* — actions and inferences that the agent attributes to itself, and actions and inferences that the agent has constructed, but does not attribute to itself — that can be only done in other situations or to other people (subjects).

All the represented architectures implement the distribution of alternative scripts. These scripts are not mixed and always choose a "leader", which further controls the agent's performance at each moment. The most important difference of architecture in Figure 6 is that the choice between scripts and their evaluation, is executed by a script at the next processing level, while in the architectures depicted in Figures 2 and 3, the selection of scripts is managed by a mechanism external to the scripts space — an inhibitory process (Figure 2) or temporal distribution (Figure 3). Thus, when moving to the architecture in Figure 6, we observe an "interiorization" of the mechanism for script evaluation and selection. This cognitive evaluation, however, can select not only the most activated script of the previous level — it can take into account other, less activated and less relevant scripts. For example, it can suggest the utterance *It's a good thing you care about me!* as an ironic answer. Less activated scripts can also serve as a matter for imagination ("what could take place", "what I could do") and the theory of mind ("what another person could decide/do").

Scope and limitations of the study

Based on the example of irony, we intended to show that the processes able to explain the architecture of consciousness (demonstrate *the effects of consciousness*) operate at the boundary between the reactive and deliberative processing levels, where an r-script interacts with several activated d-scripts. As we expected, the level of processing does not play the key role here. The same effects can appear during the interaction of d-scripts: "I did M_1^f, but I feel that it is awful and I had to do M_2^f". Similar effects are possible between the deliberative processing and meta-management level, where a person evaluates their own inferences and options for action. So the effects of consciousness are connected with the way scripts interact, not with the location of the scripts in the cognitive model.

We do not claim that our computer model has simulated consciousness or at even the effects of consciousness. We rather consider the software as an illustration of the approach. We have simulated irony as a determined procedure, which

always suppresses the most activated negative script and selects a positive script with the highest activation. While moving to the computer simulation of consciousness, it is important to provide more sophisticated interaction between d- and r-scripts.

We do not claim that the random nature of the output or non-determined nature of the processor are important characteristics of consciousness. If a structural scheme of consciousness works on a determined hardware, then for a given input (stimulus S_1) and given the state of the model (scripts), the system will provide one and the same output. At the same time, the represented model has a source of pseudo-random choice: it is evident from Table 1, that at least 12 scripts have the same activation level — 2.3482. Five of these scripts can be used for an ironic answer. What is the main factor of this selection? It can be some minor factors such as the order of the scripts in the database and the sequence of their retrieval. This factor can be determined: each time, for a given stimulus $S_{1,}$ the same ironic answer will be selected for each attempt. At the same time, during the development of the model, the influence of this factor can be reduced: input structures may contain bigger sets of features — S_1 stimuli may differ, reducing the determined nature of the selection. During operation, the system may collect preferences for certain particular scripts, depending on previous choices, or, on the other hand, may avoid repetitive answers. This may appear to be a flexible reaction system, in spite of the deterministic nature of the hardware.

Conclusion

Computer systems for natural speech understanding and the simulation of irony, from our point of view, offer an illustration of an approach which can serve as a basis for further simulation of the effects of consciousness. The mechanism of mutual activation of d/r-scripts (or their analogues) and their interaction in the processing scope can be a cornerstone for the computer model — *the minimal architecture of consciousness*. Within this architecture, the agent should activate several scripts in the first stage of processing, place their if-statements or actions (inferences) within a processing scope, and differentiate the scripts according to their activation by a script of the second stage. This provides an opportunity for the agent to simulate imaginary situations, its own imaginary actions, and the pretended actions of other agents.

Acknowledgements

The research is supported by the Russian Science Foundation (RScF grant 17-78-30029) in the part of modeling semantic representation in the human brain. The design of the emotional computer agents is supported by the Russian Foundation for Basic Research (ofi-m grant 16-29-09601).

References

Bulygina, T.V., & Shmelev, A.D. (1997). *Language conceptualization of the world*. Moscow: Languages of the Slavic Cultures.

Chernigovskata, T.V. (2016). *Cheshire smile of the Schrödinger's cat: Language and consciousness.* Moscow: Languages of the Slavic Cultures.

Kotov, A. (2003). *Mechanisms of speech influence in publicistic mass media texts.* (Ph.D thesis). Russian State Humanitarian Institute, Moscow.

Kotov, A. (2007). Simulating dynamic speech behaviour for virtual agents in emotional situations. In A. Paiva, R. Prada & R. Picard (Eds.), *Affective computing and intelligent interaction* (Vol. 4738, pp. 714–715). Springer / Heidelberg, Berlin.

Kotov, A. (2009). Accounting for irony and emotional oscillation in computer architectures. *Proceedings of International Conference on Affective Computing and Intelligent Interaction ACII 2009* (pp. 506–511). Amsterdam: IEEE. doi: 10.1109/ACII.2009.5349583

Minsky, M. L. (1968). Matter, mind and models. In M. L. Minsky (ed.), *Semantic information processing* (pp. 425–431). Cambridge, MA: MIT Press.

Minsky, M.L. (1988). *The society of mind*. New York, London: Touchstone Book.

Nagel, T. (1974). What is it like to be a bat? *Philosophical Review, 83*(4), 435–450. doi: 10.2307/2183914

Nilsson, N.J. (1988). *Artificial intelligence: A new synthesis*. San Francisco: Morgan Kaufmann.

Overgaard, M., & Mogensen, J. (2017). An integrative view on consciousness and introspection. *Review of Philosophy and Psychology, 8*(1), 129–141. doi: 10.1007/s13164-016-0303-6

Searle, J. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences, 3*(3), 417–424. doi: 10.1017/S0140525X00005756

Sloman, A. (2001). Beyond shallow models of emotion. *Cognitive Processing, 2*(1), 177–198.

Sloman, A., & Chrisley, R. (2003). Virtual machines and consciousness. *Journal of Consciousness Studies, 10*(4–5), 133–172.

Valitutti, A., & Trautteur, G. (2017). Providing self-aware systems with reflexivity. *ArXiv e-prints:* 1707.08901.

Velichkovsky, B.B. (2017). Consciousness and working memory: Current trends and research perspectives. *Consciousness and Cognition, 55*, 35–45.

Velichkovsky, B.M. (2015). Soznanie [Consciousness]. *Bolshaja Rossijskaja Encyklopedia* [The Great Russian Encyclopedia] (Vol. 30, pp. 623–625). Moscow: The Great Russian Encyclopedia.

Original manuscript received August 25, 2017
Revised manuscript accepted September 11, 2017
First published online September 30, 2017