CONSCIOUSNESS IN COMPUTER ARCHITECTURES: THE THEORY OF EXPELLED SCRIPTS[1]

Artemiy Kotov

kotov@harpia.ru

Russian State University for the Humanities, Kurchatov Institute

Consciousness is a phenomenon, experienced by a person subjectively. As a naïve language concept *consciousness* may be characterized by different features, in particular: (a) it provides some space for subjective imaging, including imagination, (b) it distinguishes 'self' from 'non-Self' and simulates *the theory of mind*, (c) it may resolve conflicts and contradictions in representations by considering different options, preferring one and rejecting the others, (d) it grounds voluntary actions as deliberately selected and responsible. Consciousness has a long study history, well summarized by M. Overgaard and J. Mogensen (2017). Following B.J. Baars (1988) consciousness operates as a searchlight, highlighting concepts within the current scope of attention. Another classic view is *the multi-level approach* by D. Rosenthal, where higher-order thoughts operate on the ordinary thoughts, previously unconsciousness. Such level architectures are currently implemented in many theoretical and computer models (Minsky 1968; Sloman, Chrisley 2003). At the same time, the stipulation of several processing levels is not sufficient to capture the nature of consciousness. We start from the assumption, that consciousness is a phenomenon, appearing within some information processing architectures. While the architectures rely on classic causality, they may run not only in the human brain, but also in some artificial processors. Modern approaches to brain modelling assume, that consciousness should appear within the accurate computation model of the entire brain. However, we suggest, that some effects of consciousness may be simulated within a simplified model — *a minimal architecture of consciousness*. We try to simulate these effects withing our computer architecture for companion robots.

The architecture contains a speech processor, central processor and a robot management system, which may output speech and basic nonverbal behavior on a companion robot F-2. The central processor consists of *scripts* — if-then operators or productions. Scripts simulate inferences, speech replies and basic emotions. A script includes two model representations: M^i (initial) and M^f (final). As soon as a semantic representation (frame) of an input text matches the premise M^i, the script invokes and generates the output representation M^f, forcing the robot to perform a speech response or an emotional reaction, attached to M^f. Multiple scripts may invoke at the same time, thus generating a compound behavioral pattern on the robot. The model presently consists of 1600 *r-scripts* for rational processing of input and 79 *d-scripts* for emotional processing. In particular, the utterances *Someone is looking at you* or *a man is always interested in the life of the beloved girl* invoke d-scripts ATTENTION (*Everybody appreciates me!*), PLAN (*They plan something against me!*) and SUBJV (*He always does one and the same, nothing else! He always thinks about one and the same!*). The performance of scripts, tested within processing of 80 mln. wordform text corpus, suggests an important view on the essence of consciousness. We shall concentrate on the ability of scripts to classify an input situation into several classes at the same time.

As shown in Fig. 1a, a simple reactive system compares an input S_1 to the premises (M^i) of all the available scripts. The best script (here — *d-scr1*) is selected for processing, while the alternative scripts (like *d-scr2*) are suppressed. It means, that conceptualization M_1^i, suggested by the best script *d-scr1* will be the only processing route, like, for the example above: *everybody appreciates me!* (ATTENTION). In this case the processing scope of attention ("searchlight") will include, first, the initial model M_1^i of *d-scr1* (see scope A), and then — its final model M_1^f (see scope B).

[1] The present study has been supported by the Russian Science Foundation, project No 19-18-00547.

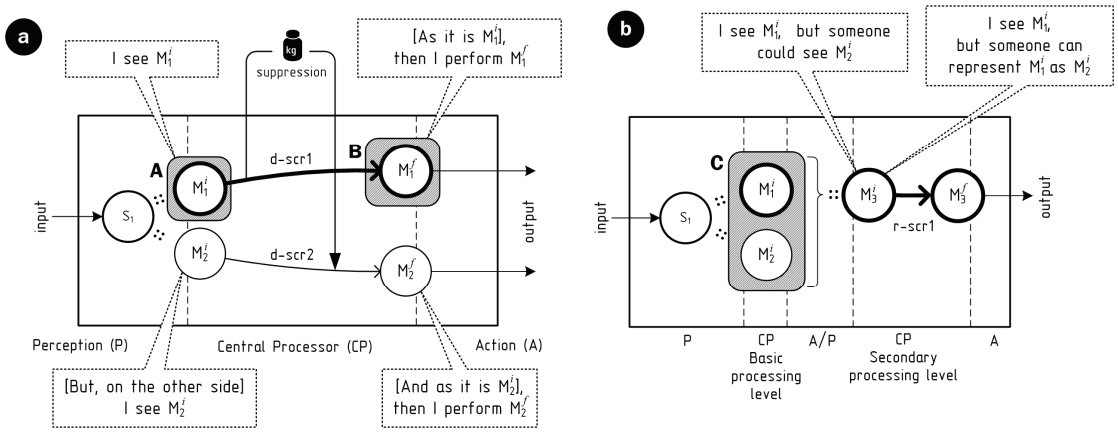

Fig. 1. (a) Simple architecture: the leading script *d-scr1* suppresses other scripts, like *d-scr2*, (b) The minimal architecture of consciousness: script *r-scr1* of the secondary *deliberative* level gains access to the diverse representations M_1^i, M_2^i, constructed by scripts (not shown) on the basic processing level

Within the simple architecture, the alternative script *d-scr2* may not be completely suppressed, but instead delayed in time. The system will process *d-scr1*, represent S_1 as a positive situation (via ATTENTION), then return to *d-scr2*, as to the second activated script, and represent S_1 as a negative situation (via PLAN and SUBJV scripts). In this case the system processes scripts separately from each other and combines their cues within the behavior. Although a system, which selects only one, the most relevant script, could be quite effective, secondary or *expelled* scripts (like *d-scr2*) may be useful to give an alternative view on an incoming situation S_1. Within this architecture we have simulated irony: the robot received an event 'somebody hits you', invoked DANGER script as the most relevant, suppressed its expression and searched for the best positive d-script for an ironic answer. The result could be the ATTENTION with an output *It's good you paid attention at me!* For the simulation of irony, a separate *irony script* has been checking the list of activated scripts (scope C on Fig. 1b), looking for an alternative script in case the relevant script is suppressed. We assume, that within this mechanism a higher-order rational script *r-scr1* (Fig. 1b) should check the most relevant primary script, identify M_1^i as the most relevant ('self') representation of S_1, and use an alternative ('non-Self') representation M_2^i for (a) irony, (b) self/non-Self distinction (M_1^i is 'my view' and M_2^i is 'not my view'), (c) for imagination and the theory of mind ('in another situation I would represent S_1 as M_2^i' or 'someone else could represent S_1 as M_2^i'). Thus, we suggest, that within the model of alternative representations, the expelled scripts may serve as a basis to simulate irony, imagination, theory of mind, conflict handling and other features of consciousness, suggested at the beginning of this note, thus, *representing a minimal architecture of consciousness*.

References

Overgaard M., Mogensen J. 2017. An integrative view on consciousness and introspection // Rev. Philos. Psychol. Vol. 8. No 1, pp. 129–141.

Baars B. J. 1988. A cognitive theory of consciousness.

Rosenthal D. 2005. Consciousness and mind.

Sloman A., Chrisley R. 2003. Virtual Machines and Consciousness // J. Conscious. Stud. Vol. 10. No 4–5, pp. 133–172.

Minsky M. 1968. Matter, Mind and Models // Semantic information processing. Cambridge, Mass.: MIT Press, pp. 425–431.