ARS: An AGI Agent Architecture

Samer Schaat, Alexander Wendt, Matthias Jakubec, Friedrich Gelbard, Lukas Herret, and Dietmar Dietrich

Institute of Computer Technology, Vienna University of Technology, A-1040 Vienna
{schaat,wendt,jakubec,gelbard,herret,dietrich}@ict.tuwien.ac.at

Abstract. The computational paradigm in Cognitive Science, which the AGI approach revives, provides a powerful methodology of examining human information processing by testing assumptions in computer simulations, and enables technical applications with human-like capabilities. Nevertheless, intensive interdisciplinary collaboration and the development of a holistic and integrated model remain ongoing challenges. This includes the consideration of the basis of rational cognition, in particular the significance of unconscious and affective processes in the human mind. We take these issues into consideration and integrate them into a holistic and integrated functional model of the human mind, implemented as an agent's decision unit and evaluated in an Artificial Life simulation using an interdisciplinary methodology.

Keywords: Artificial General Intelligence · Cognitive Architectures · Computational Simulation · Artificial Recognition System · Artificial Agents.

1 Introduction

AGI (Artificial General Intelligence) "as the science of the mind as a computational system" [1], has revived the original endeavour of Cognitive Science and Artificial Intelligence to find a unified description of cognition instead of solving specific problems as in current conventional AI. Examining cognition using a synthetic approach is a powerful way to understand the human mind and enables us to test our ideas of how the mind works by running them as a computer simulation. Furthermore, we can use this knowledge to develop technical systems with human-like capabilities; for in an engineering sense, building a system and understanding it go hand in hand.

Nevertheless, two key aspects are often neglected in AGI models: (1) serious and regular interdisciplinary cooperation between the different disciplines concerned with studying the mind, and (2) taking into account the relevance of processing principles of the unconscious, especially affective processes.

Regarding the first aspect, in an interdisciplinary collaboration, computer science provides powerful techniques for developing and testing a deterministic model of the human mind by using approaches from information theory such as computer simulations, layered models, separation and modelling of data and functions, top-down design processes, and requirements engineering; at the same time neurobiology and psychology provide insights into the mind. Such a collaboration requires an

B. Goertzel et al. (Eds.): AGI 2014, LNAI 8598, pp. 155–164, 2014.

interdisciplinary methodology. With regard to the second aspect, traditional cognitive architectures focus on modelling rational thinking. Such approaches often underestimate the significance of the unconscious, whose key role is emphasized by many disciplines (e.g. [2, 3]). Hence, a holistic model of human decision making must consider the unconscious foundations of rational cognition and integrate them with models of rational thinking into a unitary model.

2 Related Work

The ARS (Artificial Recognition System) presented in this paper possesses attributes to be classified as following a cognitivist approach.

ACT-R (Adaptive Control of Thought-Rational) models an integrated theory of the human mind and consists of several encapsulated modules. Their functionalities are mapped to the cortical regions in the brain [6]. ACT-R is based on the multi-store model theory [7] and therefore implements different memory systems and operations for each module.

SOAR (State, Operator Apply Result) is a realization of the two hypotheses of classical artificial intelligence by Newell and Simon [8, 4, 5]: "The physical symbol system hypothesis" and the "heuristic search hypothesis". They state that such a system "…has the necessary and sufficient means for general intelligent action" and that a solution will be found "by generating and progressively modifying symbol structures…" [8]. In contrast to ARS, SOAR is not based on human ways of thinking [9]. A difference to ACT-R is that the production rules in SOAR are relative simple, i.e. they only execute one change of the working memory. They are fired in parallel, while in ACT-R productions may be extensive as one rule can alter many buffers [10].

Differing from the previously described architectures, BDI (Belief, Desire, Intention) architecture as described in [12] is not a problem solving system based on a heuristic search. It is based on a theory of practical reasoning, which is a planning theory of intention according to [11]. Its foundations are the three mentalistic attributes belief, desire and intention, which define the state of the agent. Beliefs describe the agent's view of the world. Desires represent the long-term goals. They are activated by certain beliefs. Intentions are high-level plans which may be executed in order to satisfy a desire. Desires are activated depending on the activated beliefs. BDI does not use a heuristic search like SOAR and ACT-R to find a solution, but rather applies a case-based approach [13].

LIDA (Learning Intelligent Distribution Agent) is based on a combination of recent theories of the mind which are merged into a single cognitive architecture. Among them is Global Workspace Theory, which is a connectionist theory and the most widely accepted psychological and neurobiological theory of the role of consciousness in cognition [14, 16]. It is also a realization of the H-CogAff architecture [15]. Due to the connectionist approach of Global Workspace Theory, the sensors demand embodiment for the agent [16], a factor not required for the previously described architectures. LIDA uses concepts like emotions and feelings for the evaluation of situations. In contrast to the other architectures, it also defines a preconscious

and a conscious part of the system, where data is pre-processed, and through an attention mechanism a subset of the data is consciously broadcast to activate possible options for action [17].

3 ARS Approach and Model Overview

When work began on ARS the original idea was to design an intelligent system capable of recognizing and understanding real-world situations, e.g. potentially dangerous situations such as easily accessible knives threatening children in the kitchen or similar scenarios [18]. Soon it became clear that human beings can perform this type of recognition tasks because they possess something we call "feeling"[1], a feeling for the situation they observe, a feeling for the use that may be made of available objects, a feeling for how they should assess the characters and moods of others; and it became obvious that it would be anything but simple to create an artificial system with this kind of ability. What was required was nothing less than to design a model of what is called the human psyche, the psychic or mental apparatus [19], and thus design an AGI architecture. We understand the psychic apparatus as the control unit of the human organism. It is built from the nervous system with its main part – the brain –, but to understand its workings it must be described on a higher abstraction level than just the function of the neurons. If we as technicians want to build a model of the psyche at this level, we cannot determine the corresponding functions by ourselves. We need a consistent holistic functional psychic theory, and the only truly adequate theory we have been able to find is psychoanalytic metapsychology. So the ARS projects aims at concretizing metapsychology into a technical model of the human mind [18].

Fig. 1 shows the ARS model at the track level. Each track is built from several function modules. The psyche of different individuals never differs in regard to these functions, respectively their algorithms, but only in regard to data such as personal parameters or memory contents. The ARS model identifies four different input tracks: environment perception and body perception which together form the perception track, and self-preservation drives and sexual drives which flow into the drive track. This input signals the current needs of the organism. It is cathected[2] with certain psychic intensity, i.e. it is prioritized with a measure of its present importance, and may be associated with memory traces of the means to meet the respective needs. These associations again are of different psychic intensity, and an adequate reaction could be selected based exclusively on them. So far, this functionality corresponds to what psychoanalysis terms the Id, which is rather animalistic. But the humanoid agent is a social creature. Growing up, it encounters a number of rules describing desirable behaviour as a member of the group, psychoanalytically known as the Super-ego. In the defence track, conflicts between the needs of the Id and the commandments of the Super-ego are decided by functions of the third major functional block: the Ego. All this work by the psyche is entirely based on the so-called pleasure principle, and remains unconscious to people; psychoanalysis calls it the primary process. But

[1] Here "Feeling" is used as an everyday English word. Later in the project we defined it based on Damasio's theory (see chapter 4).

[2] Cathexis describes the attribution of quota of affect to psychic content. As a result this content is valuated.

evolution has equipped humans with an additional mechanism to control the unconscious output through rationally based decision making.

In the transformation track the contents resulting from the primary process are connected with word-presentations. So they become preconscious.

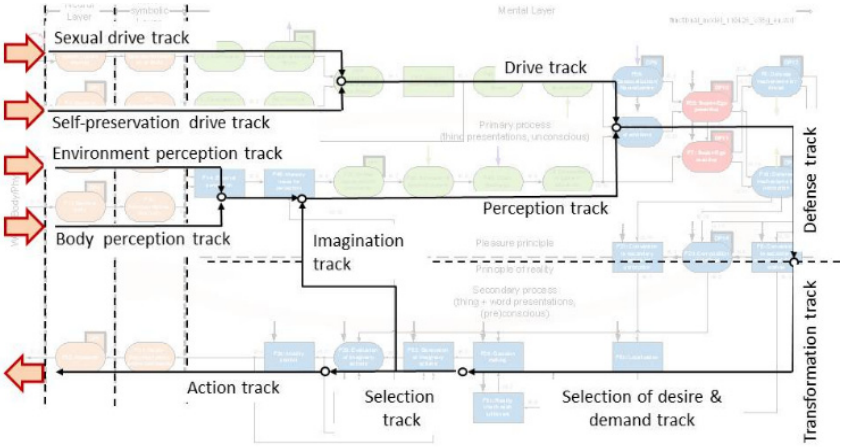

Fig. 1. ARS Model at the track level

In case of hypercathexis (with extra psychic intensity) they become conscious. The so-called secondary process includes the ability to deal with order, time, sequence and language-based models of the world as well as logical reasoning. It enables the agent to withdraw the purely pleasure-driven actions and follow the reality principle instead. It gives the agent the feeling of free will, of agency, the possibility not to act [19]. In the selection of desire & demand track, decisions are made as to which demands should be checked for possible satisfaction. The selection track finally decides which action plan to fulfil, and instructs the action track to realize it. Imagined results are fed back to the primary process by the imagination track and thus become perceived fantasy.

4 Motivations and Valuations

A fundamental question in AGI agents concerns the source for the agent's agenda and how the agent may cope with the external world while pursuing this agenda. Using the drive concept of Freud as a framework and concretizing it by Damasio's model [3] of emotions, we use a generative multi-level model of motivations and valuations to tackle these questions. Based on bodily needs, valuations generate and prioritize motivations (drive representations) which are transformed into goals. These valuations occur incrementally following different principles and influences.

As shown in Fig. 2, organic tension values from the agent's body are represented as psychic intensity in the psychic layer. In the process of generating drive representations, psychic intensity is represented as a quota of affect and used to valuate memorized

objects and actions according to the pleasure principle, which valuates that content as the best which brought the most satisfaction in the past. This valuation may however be changed by defence mechanisms (see chapter 6). The next valuation step uses neutralized intensity, which is a personality-specific part of the drives' quota of affect, to extend the valuation of memorized content according to the reality principle, i.e. the consideration of affordance in the environment (see chapter 7).

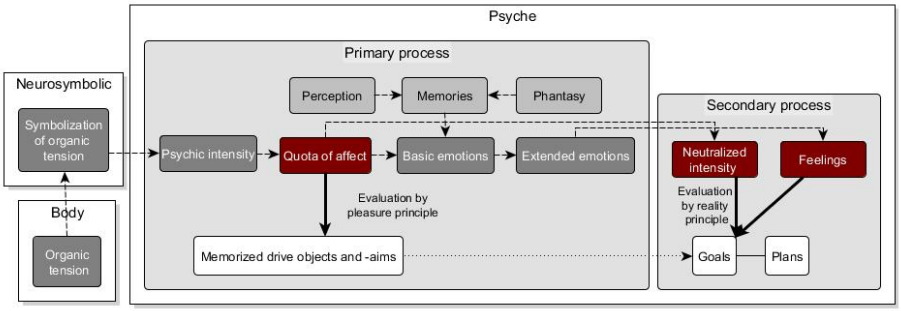

Fig. 2. An incremental multi-level model of valuations

In general, logic and time issues are considered in this valuation step, which transforms valuated drive representations to prioritized goals. The valuation of goals can be extended by feelings as conscious representations of emotions, which are generated based on all quotas of affect, and memorized emotions that are activated by perception and phantasy.

5 Perception

Perception is modelled as a means for the fulfilment of the agent's motivations. In this regard perception supports matching valuated memories with objects in the external environment. This results in constructing images which include all sensual modalities, and provides the information on how to fulfil the agent's motivations in the external world. Hence the recognition of objects is based on the agent's experience and expectations, which are generated from drives. This complies with the integration of bottom-up and top-down approaches into a holistic model of perceptual categorization, which is represented by using an activation-based exemplar model with multiple activation sources (i.e. external stimuli and expectations triggered by drives).

6 Conflict and Defence Mechanisms

One of the challenges in AGI systems are conflicts in decision making. In the ARS project, conflicts arise in the following cases: differences between drive wishes of the agent, the possible fulfilment of those drive wishes in the simulation environment, emotions, and social rules of the software agent. We implement psychoanalytic defence mechanisms to resolve these conflicts and therefore to filter and/or alter input

data of the software agent. The defence mechanisms under consideration are repression, denial, reaction formation, reversal of affect, displacement, idealization, and depreciation [20]. Fig. 3 sketches the functionalities of the defence mechanisms. First we must investigate how conflicts in AGI systems can be detected and assessed. For this purpose we implement two modules which represent the Super-ego. The Super-ego modules compare drive wishes of the agent, emotions of the agent, perception, and social rules for conflicts.

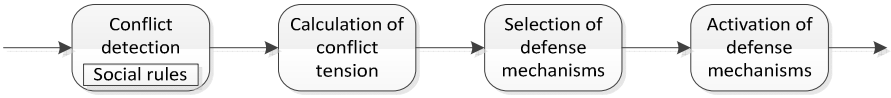

Fig. 3. Functionalities of defence mechanisms

If a conflict is detected, the weights – or rather the quotas of affect – of the conflicting components are summed up to obtain the value of the conflict tension. The detection of conflicts is implemented by the use of rules: if a match of the left side of a rule is found, the right side of that rule indicates which drive wish, emotion, and/or perception is to be defended.

Once a conflict has been detected and its conflict tension calculated, the agent must decide which defence mechanisms to select and activate. Hence the defence mechanisms are sorted, from primitive to high-level defence mechanisms. The basic factor for selecting a defence mechanism is the ego strength, which is represented by the sum of available neutralized intensity. According to the current ego strength a defence mechanism is selected and activated.

7 Decision Making and Planning

Decision making in ARS is a deliberative, two-stage selection process. In contrast to the unconscious and rather more reactive parts of the system, the processing of a goal requires several model cycles. The decision making and planning process is illustrated in Fig. 4. The first step is to extract goals in 1a), 1b), and 1c); a goal is a container which consists of the goal type, a goal object, plans, importance, and status flags. In 1c), motivations for what the agent shall achieve are extracted from drives. A drive, which originates from the homeostatic needs of the body, is converted to a goal called the aim of the drive. Such a drive may be e.g. to satisfy the need to eat. For decision making, the current state of the feelings is also used in the evaluation of situations in 1b). All options or possible goals available for the selection are extracted from two sources in 1a): either directly from the perception or from the activated memories. A possible goal, which is extracted from a certain drive representation, tells the agent that the perceived object would satisfy the need to eat if selected. The activated memories have the structure of sequences of events and are assembled from independent, activated events as images in the unconscious part of the system. These memories are the beliefs of the system, as they are sequences of actions and consequences in different situations. For instance, such a sequence might inform the agent of a dangerous situation

and potential consequences by creating a possible goal from a bad feeling which is associated with that sequence. In 2a), all incoming goals receive initial status flags as a result of a basic evaluation of the effort and whether they are reachable for the agent or not. The status flags of the goals are also used for teleo-reactive planning [21]. 2b) will be explained later as it applies only to the subsequent model cycle in the multi-cycle decision process.

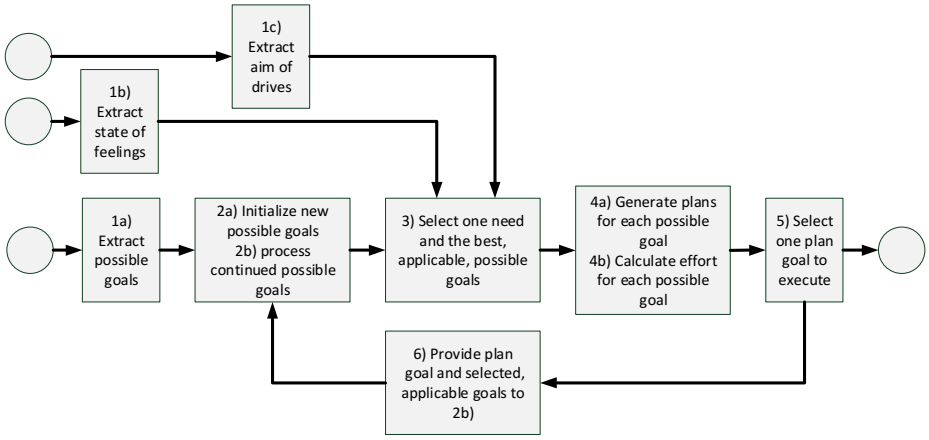

Fig. 4. Process model of decision making and planning in ARS

In 3), the first stage of decision making is completed: all possible goals are evaluated based on their likelihood of fulfilling the incoming aims of drives under consideration of feelings. As a result, one or more possible goals are selected which have the ability to satisfy the strongest aims of drives. For those possible goals which are relevant to fulfilling the most urgent needs, or which demand a reaction to a situation, plans are generated in 4a) and the effort of executing those plans is estimated in 4b). Finally in 5), the goal with the highest importance, i.e. the best possibility of satisfying a need with the lowest effort or the avoidance of harm to the body, is selected in the second stage of decision making. It is called the plan goal. All goals from the first stage of decision making are then stored in short-term memory.

The action plans attached to the plan goal are either external or internal actions. In the case of an external action, the action command is sent to the body for execution. In the case of an internal action, an internal action command is executed within the model in the next cycle. For each executed external action, several internal actions are usually executed first, e.g. to focus on an object before starting to move towards it.

In the next model cycle, the stored possible goals – include the plan goal in the short-term memory – are continued in step 6) of Fig. 4 as they are compared and merged with newly extracted possible goals. In 2b), enhanced analysis of continued goals is performed, triggered by internal actions. As a result, new status flags are defined and associated with the possible goal. These status flags influence the evaluation of the possible goal. Then in 3), all possible goals are evaluated again. As each goal is handled independently, the agent is able to continually consider new goals and situations and to pause the pursuit of the current plan goal.

8 Simulation Architecture

The implementation of ARS runs within an artificial life simulation, which is based on the multi-agent simulation framework MASON. However, because of the generic interface of the ARS architecture, it is also possible to use it in other applications. The MASON framework provides a scheduler, a physics engine, a control panel and visualization tools. The execution of each simulation object is divided into three parts: sensing, processing, and execution. The simulation cycle is also executed for each part of all simulation objects, i.e. first all sensing parts of all objects are executed, then the processing parts, and finally the execution parts.

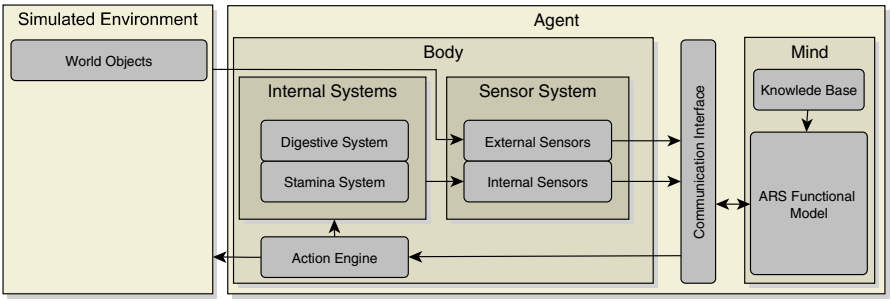

Fig. 5. ARS implementation architecture

As seen in Fig. 5, the agents are composed of a body and a mind, i.e. the ARS cognitive architecture. The body, connected with the ARS architecture, consists of internal systems like a digestive system (energy balance), body internal sensors like blood sugar, external sensors like vision, and an action engine which executes action commands from the mind. The mind contains the implementation of ARS, together with the knowledge base which is based on Protégé Frames.

9 Evaluation

As introductory mentioned, the development of an interdisciplinary methodology is a key challenge when developing and evaluating AGI agents. We use a case-driven methodology that guides interdisciplinary cooperation. This means that psychoanalysts and neuroscientists use their experience of real-world conditions to write an exemplary case describing a situation (e.g. a hungry agent). This case is structured in a simulation case which allows analysis of the required functions and data for the simulation model. An overview of such a simulation case is provided in Fig. 6. To evaluate the model, agent-based simulation is harnessed, which enables testing of our assumptions and the plausibility of the model. In particular, we validate whether the agents behave as expected (i.e. as described in the simulation cases). This includes observing whether changing the data results in the expected behaviour (as described in the simulation case). Since in this overview article no room for simulation results exist, see e.g. [22] for details.

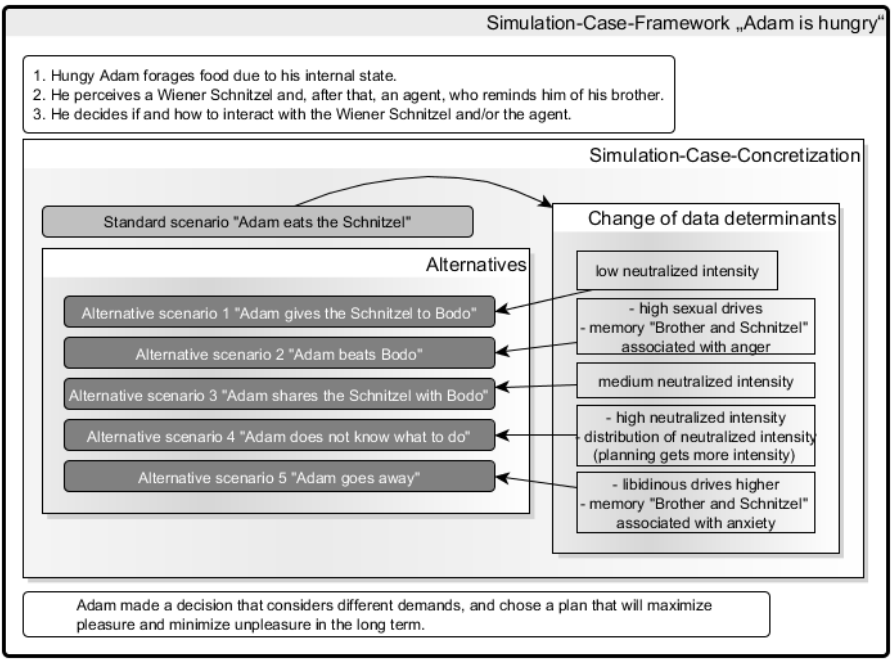

Fig. 6. Simulation case for the description of the behaviour of two agents, Adam and Bodo

10 Conclusion

We have shown how processes that follow the principles of the unconscious can be integrated with rational aspects of decision making to approach how a decision unit for an AGI agent that mimics human information processing may be developed. In particular, we consider affective processes for the valuation of data – in keeping with basic principles of rational cognition – and defense mechanisms for handling conflicting trains of thought within the agent. We integrate these functions with perception, rational decision making and planning, and evaluate the holistic model in an Artificial Life simulation using an interdisciplinary methodology. With regard to future work, probably the most notable shortcoming of the current model is the agent's limited ability to learn.

References

1. Bach, J.: A motivational system for cognitive AI. In: Schmidhuber, J., Thórisson, K.R., Looks, M. (eds.) AGI 2011. LNCS, vol. 6830, pp. 232–242. Springer, Heidelberg (2011)
2. Bargh, J.A., Chartrand, T.L.: The unbearable automaticity of being. American Psychologist 54(7), 462–479 (1999)
3. Damasio, A.: Looking for Spinoza: Joy, Sorrow, and the Feeling Brain. Harvest Books, Washington (2003)

4. Vernon, D., Metta, G., Sandini, G.: A survey of artificial cognitive systems: Implications for the autonomous development of mental capabilities in computational agents. IEEE Transactions on Evolutionary Computation 11, 151–180 (2007)
5. Langley, P., Laird, J.E., Rogers, S.: Cognitive architectures: Research issues and challenges. Cognitive Systems Research 10(2), 141–160 (2009)
6. Anderson, J.R., Bothell, D., Byrne, M.D., Douglass, S., Lebiere, C., Qin, Y.: An integrated theory of the mind. Psychological Review 111(4), 1036–1060 (2004)
7. Atkinson, R.C., Shiffrin, R.M.: Human memory: A proposed system and its control processes. In: The Psychology of Learning and Motivation, vol. 2, pp. 89–195. Academic Press, New York (1968)
8. Newell, A., Simon, H.A.: Computer science as empirical inquiry: symbols and search. Commun. ACM 19(3), 113–126 (1976)
9. Newell, A.: Unified Theories of Cognition. Harvard Univ. Press (1994)
10. Turnbull, D.G., Chewar, C.M., McCrickard, D.S.: Are cognitive architectures mature enough to evaluate notification systems? In: 2003 International Conference on Software Engineering Research and Practice (SERP 2003), Las Vegas NV (2003)
11. Bratman, M.: Intention, Plans, and Practical Reason. Harvard University Press (1987)
12. Gottifredi, S., Tucaty, M., Corbatta, D., Garcia, A.J., Simari, G.R.: A BDI architecture for high level robot deliberation. Inteligencia Artificial 46, 74–83 (2010)
13. Wendler, J., Hannebauer, M., Burkhard, H.-D., Myritz, H., Sander, G., Meinert, T.: BDI design principles and cooperative implementation in roboCup. In: Veloso, M.M., Pagello, E., Kitano, H. (eds.) RoboCup 1999. LNCS (LNAI), vol. 1856, pp. 531–541. Springer, Heidelberg (2000)
14. Shanahan, M., Baars, B.: Applying global workspace theory to the frame problem. Cognition 98(2), 157–176 (2005)
15. Sloman, A., Chrisley, R.: More things than are dreamt of in your biology: Information-processing in biologically inspired robots. Cognitive Systems Reasearch 6(2), 145–174 (2005)
16. Faghihi, U., Franklin, S.: The lida model as a foundational architecture for agi. In: Theoretical Foundations of Artificial General Intelligence, pp. 103–121. Springer (2012)
17. Franklin, S., Ramamurthy, U.: Motivations, values and emotions: 3 sides of the same coin. In: Proceedings of the Sixth International Workshop on Epigenetic Robotics, Paris, France, vol. (128), pp. 41–48. Lund University Cognitive Studies (September 2006)
18. Solms, M.: What is the 'Mind'? A Neuro-Psychoanalytical Approach. In: Dietrich, D., Fodor, G., Zucker, G., Bruckner, D. (eds.) Simulating the Mind, pp. 115–122. Springer, Vienna (2009)
19. Dietrich, D., Fodor, G., Zucker, G., Bruckner, D.: Simulating the Mind A Technical Neuropsychoanalytical Approach. In: Proceedings of the 1st ENF - Emulating the Mind, 2007 Conference, Vienna (2009)
20. Eagle, M.N.: From Classical to Contemporary Psychoanalysis, A Critique and Integration. Routledge, Taylor and Francis Group, LLC (2011)
21. Nilsson, N.: Teleo-reactive programs for agent control. arXiv preprint cs/9401101 (1994)
22. Schaat, S., Doblhammer, K., Wendt, A., Gelbard, F., Herret, L., Bruckner, D.: A Psychoanalytically-Inspired Motivational and Emotional System for Autonomous Agents. In: Proceedings of the 39th Annual Conference of the IEEE Industrial Electronics Society, pp. 6648–6653 (2013)