Volitron: On a Psychodynamic Robot and Its Four Realities

Andrzej Buller
ATR Human Information Science Laboratories
2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288 Japan
buller@atr.co.jp

Abstract

This paper discusses the concept of Volitron—a controller to make its host robot increase its competence in such activities as self-initiated exploration of an environment, new goal acquisition, and planning/executing of actions while taking into account predicted behaviors of objects of interest. There are four key elements in Volitron's structure: a model of perceived reality, a model of desired reality, a model of ideal reality and a model of anticipated reality. The task of a robot's working memory includes producing images of the robot itself imitating another subject's activities and sending the images to a model of desired reality. A tension (a concept borrowed from psychoanalysis) arising from the differences between a perceived reality and a desired reality is a source of a motivation toward action. The final decision to take an action is based on a comparison of the model of anticipated reality with that of ideal reality. The interaction of Volitron's elements are described in the paper. Furthermore, a computational model of working memory (WM) and its psychological justification are provided.

1. Introduction

Volitron is to give its host robot an increasing competence in such activities as searching for sources of energy, friend/enemy recognition, self-initiated exploration of an environment, social cooperation, acting in accordance with acquired principles, acquiring new goals, and planning sequences of actions while taking into account predicted behaviors of objects of interest (people, animals, other robots, plants, inanimate devices). As for the source of the robot's motivation to do anything, a kind of psychic *tension*—a key concept in psychoanalytic theory—has been adopted.

According to psychoanalytic theory, the course taken by mental events is invariably set in motion by unpleasant tension, and this course takes a direction leading to a lowering of that tension, which means the avoidance of unpleasantness or the production of a pleasure (Freud 1920/1990:3). For Volitron, it is assumed that tension is one of the variables characterizing a state of the working memory (WM).

Folowing Tulving's suggestion about the structure of the human memory system (Tulving 1995), Volitron is based on five kinds of memories: PRS (perceptual representation system), procedural memory, semantic memory, episodic memory, and working memory (WM). While the first four types of memories are used to store information for a relatively long time, WM is primarily for processing pieces of information taken from the other kinds of memories and for changing the contents of these memories. Buller (2000) proposed that the pieces of information are called *memes* and that they occur in WM in multiple copies. A population of *memes of satisfaction*, as well as a population of *memes of dissatisfaction*, are generated as by-products of the WM's activity. Based on numerical balance of these two populations, a level of tension is calculated. Volitron is hard-wired to continuously attempt to maintain a level of tension that is as low as possible.

Where do the memes of satisfaction /dissatisfaction come from? According to Freud's psychodynamic theory, mental life is a kind of a continuous battle between conflicting psychological forces such as wishes, fears, and intentions. In artificial minds, the psychological forces can be understood as changing the allocation of available computational resources. For the Volitron-based 'mind' it is proposed that populations of contradictory memes remain in continuous battle, attempting to expel "enemy forces" from WM and thus occupy as many of the WM's computational power as possible. The difference between a perceived reality and a desired reality is proposed as the primary reason for the appearance of the memes of dissatisfaction. Technically speaking, each of the realities is a knowledge base constituting a specific model of the world. In consideration of an intrinsic dynamics of meme populations and the essential role of dynamic processes in Freud's theory, let Volitron-controlled robots be called *psychodynamic robots*.

2. Volitron architecture and embeded functions

The proposed paradigm of Volitron's learning and self-development is model-based. As Figure 1 shows, four separate models are employed: (1) a model of perceived reality, (2) a model of desired reality, (3) a model of ideal reality, and (4) a model of anticipated reality.

All acquired knowledge of perceived and desired reality is stored in the semantic memory which is static. Upon request, selected pieces of knowledge, called memes, are injected from the "storehouse" of semantic memory into the "theater" called working memory (WM). Memes occur in WM in multiple copies and interact with one another, as well as with the working memory itself. Several varieties of meme processing takes place in the working memory. The most essential are: (i) categorization, (ii) hunger-for-knowledge production, (iii) imitation-drive production, (iv) domination-drive production, (v) model comparison, (vi) action-drive production, (vii) candidate-plan generation, (viii) anticipated-reality creation, (ix) judgment of candidate plans and anticipated results of implementation of the plans, and (x) defense mechanisms. Below the roles and mutual relations of the processes are suggested.

Categorization involves interactions between memes coming from PRS and memes provided by the model of perceived reality. This process results in the production of memes representing candidate recognitions. Contradictory memes annihilate each other, hence a particular meme population may dominate in WM, which may cause a modification of the model of perceived reality.

Hunger-for-knowledge production is caused by the memes that carry the knowledge of missing elements in the model of perceived reality. A stream of such memes activates the device for *action-drive production* in an attempt to gain the missing elements.

Imitation-drive production is an operation conducted on memes that represent a perceived action performed by observed people or other robots. Those memes are transformed into memes that represent the same action as fictitiously performed by the robot itself. The new memes are directed to the model of desired reality. Analogously, memes representing a perceived object are transformed by the *domination-drive production* unit into memes representinging the object as a property or a subordinate of the robot.

Model comparison applies to the model of perceived reality and to the model of desired reality. Memes coming from the models are processed, which results in the production of memes of dissatisfaction. The level of tension is calculated based on the number of memes of dissatisfaction vs. the number of memes of satisfaction. The tension resulting from this comparison activates the *action-drive production* toward a physical action using actuators. The goal of this physical action is to change the environment in such a way that the perceived reality becomes more similar to the desired reality.

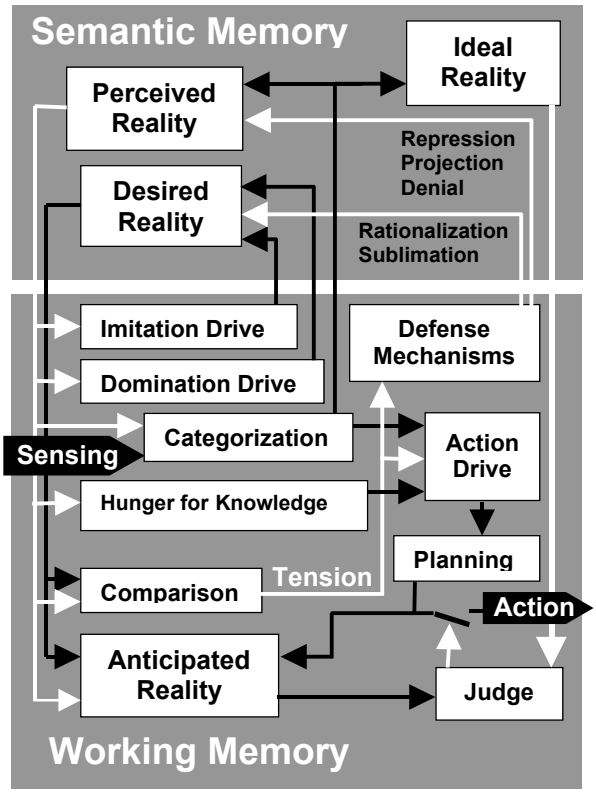

Figure 1. Structure of Volitron

Candidate-plan generation is a continuous process stimulated by the drive to action. Multiple plans are generated and tested. The planning may be supported by data coming form the model of perceived reality. Testing of a candidate plan involves simulating changes in a portion of the model of perceived reality based on a fictitious execution of the plan. In this way, a virtual anticipated reality appears in the working memory.

Judgment of candidate plans and anticipated results concerns the process of preventing Volitron from executing a lethally stupid action. A device called "Judge" compares the anticipated reality with the relevant part of the model of ideal reality and, depending on the result of the comparison, allows an acceptable plan to be physically executed.

Defense mechanisms start working when no satisfactory plans has been generated for a long time or when the already executed plan could not change the environment so as to reduce the tension. Memes produced by the defense mechanisms can cause changes in the models of reality. As for the model of perceived reality, the changes may manifest themselves as a repression of inconvenient facts, a projection of one's failure onto another object (a person, an animal, or another robot), or a denial (assumed to be an annihilation of certain types of memes coming from the sensory system). As for the model of desired reality, the changes may manifest

themselves as a rationalization (a justification for removal of unrealistic desires) or a sublimation (a conversion of unrealistic desires into cues for a substitute activity). Owing to these changes, the tension will eventually be reduced in some way. The price of achieving tension reduction inn this way is an inadequate model of perceived reality. Nevertheless, owing to the tension-driven mentality, a psychodynamic robot will be strongly motivated to do anything and act on its own.

3. Working Memory (WM)

A prototype WM for Volitron has been simulated as a grid of channels connecting processing units, called *tiles*, dedicated to meme interaction. Every tile has three input channels and three output channels. Memes are mobile entities moving continuously all over the grid and interacting with one another when they simultaneously enter the same tile. As for meme structure, it was assumed that every meme represents an ordered pair of elements, where the first element is a proposition and the second element is either ∅ or a proposition. When (a) the second element of a meme is ∅, (b) the meme is a member of a population of identical copies inhabiting WM, and (c) the population dominates in WM for a certain period of time, then the first element of the meme is interpreted as a temporary belief of an agent equipped with the WM.

Four kinds of meme interactions has been implemented: *annihilation*, *positive cross-over*, *negative cross-over*, and *collision*. Annihilation takes place when memes ⟨p | ∅⟩ and ⟨q | ∅⟩ encounter each other p and q are contradictory propositions. Positive cross-over takes place when meme ⟨p | ∅⟩ encounters meme ⟨q | p⟩. In such a case, ⟨p | ∅⟩ becomes ⟨p | p⟩ and is annihilated immediately, while ⟨q | p⟩ becomes ⟨q | ∅⟩. Negative cross-over takes place when meme ⟨p | ∅⟩ encounters meme ⟨q | r⟩ and p contradicts r. In such a case, ⟨p | ∅⟩ becomes ⟨p | r⟩ and is annihilated immediately, while ⟨q | r⟩ becomes ⟨s | ∅⟩ such that s contradicts q. Collision causes a change in the directions of the memes' movement in a way analogous to a collision of balls on a snooker table. The current implementation of WM is a traditional simulation model. The next implementation is being prepared as special neural networks (Buller et al. 2002a) to be run on dedicated hardware.

In the framework of one of the experiments, streams of memes that were injected into WM represent the vagueness of social perception. The uncertainty of whether a perceived person is 'nice' from the point of view of the simulated subject is represented by streams of contradictory memes: ⟨'nice' | ∅⟩ and ⟨'not nice' | ∅⟩. The uncertainty of whether the perceived person is 'rich' from the point of view of the subject is represented by streams of contradictory memes: ⟨'rich' | ∅⟩ and ⟨'not rich' | ∅⟩. The subject's criteria of attractiveness are represented by the streams of memes: ⟨'[I can] agree [to a date with somebody who is]' | 'nice'⟩ and ⟨'[I can] agree [to a date with somebody who is]' | 'rich'⟩. For non-ambiguous data, i.e. when the stream of memes ⟨'nice' | ∅⟩ is much denser than the stream of memes ⟨'not nice' | ∅⟩ and the stream of memes ⟨'rich' | ∅⟩ is much denser than the stream of memes ⟨'not rich' | ∅⟩, WM quickly becomes dominated by the population of memes ⟨'agree' | ∅⟩. For ambiguous data, i.e. when densities of the streams of memes ⟨'nice' | ∅⟩ vs. ⟨'not nice' | ∅⟩ and ⟨'rich' | ∅⟩ vs. ⟨'not rich' | ∅⟩ are comparable, the result is counterintuitive—the populations of ⟨'agree' | ∅⟩ and ⟨'not agree' | ∅⟩ took turns in dominating in WM (Figure 2). In other words, the state of WM goes to a 2-state limit-cycle attractor called "strange attractor".

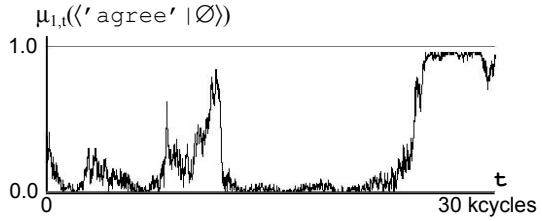

Figure 2. Sample plot of meme population dynamics in working memory (WM). $\mu_{1,t}(\langle$'agree' | ∅$\rangle)$ reflects the numerical balance of memes ⟨'agree' | ∅⟩ vs. memes ⟨'not agree' | ∅⟩.

In experiments it was shown that the simulated subject hesitated in deciding whether to accept the proposal to have a date with the person of interest. This hesitation is based exclusively on the intrinsic dynamic properties of WM fed with streams of memes, even those having constant densities (Buller 2000). Hence the presented WM has strong psychological justification, since the experiments with human subjects of Nowak and Vallacher (1998) showed that people's judgment may oscillate between highly positive and highly negative values even in the absence of new data about a person of interest. Hence, a robot's hesitation whether to continue its mission or to search for a battery charger may be expected.

Another psychological justification for the model comes from the results of an experiment in which the final configurations of meme population were interpreted as classifications of simple patterns after a single exposure of the system to one representative of each of three classes of interest (Buller et al. 2002b).

4. Discussion

Even in the newest textbooks on artificial intelligence, a 'world model' is assumed to be an indispensable part of

an intelligent system (e.g. Albus & Meystel 2001: 196). However, the idea of model-based robot intelligence has been undermined by the followers of 'intelligence without representation' (see Brooks 1991). Brooks et al. (1998) collected psychological evidence supporting the theses that humans do not build an internal model of the entire visible scene and that there are multiple internal representations that are not mutually consistent.

Volitron does not and cannot build an internal model of an *entire* scene in its WM because the model is based on a volatile swarm of two kinds of memes: those representing only selected parts of the scene and those from the model of perceived reality. The model of perceived reality is updated (sometimes making it worse!) for its entire life, never becoming perfect, and is never accessible as a whole because of the limited capacity of WM. However, despite all the shortcomings of a world model, it seem to facilitate several useful mental mechanisms that are extremely difficult (if not impossible) to obtain by using known genetic algorithms or other machine learning methods.

A psychodynamic robot cannot prevent itself from committing stupid actions. Memes coming from the model of desired reality may interfere with the process of generation of anticipated reality, which may be interpreted as a kind of "wishful thinking". However, owing to a direct link to the action driver, a quick reaction may save the robot's life. But the same link may cause an unjustified escape or an unnecessary defensive movement, which may result in a dramatic increase in the level of tension. The stupidity of the psychodynamic robot seems to be the unavoidable price of tension-grounded autonomy and "free will".

The judgment of anticipated reality is based on the model of ideal reality based on the concepts of good and evil. According to Freud's explanation of a child's development, moral categories are acquired mostly from its parents (Freud 1923/1990: 26). A 'young' robot could be hard-wired to select an entity as a "moral authority" and follow his/her/its teachings.

In psychoanalytic terms, Volitron's model of perceived reality together with its devices for planning and judgment constitute a counterpart to the conscious part of the Ego (the defense mechanisms may be considered an unconscious part of the Ego). The model of desired reality along with the devices for drives and tense creation constitute a counterpart to the Id. The model of ideal reality can be understood as an artificial Superego (cf. Freud 1923/1990).

The most challenging issue in a building a model-based intelligence seems to be the quest for machine consciousness. Dawkins (1976/1999:59) wrote: "Perhaps consciousness arises when the brain's simulation of the world becomes so complete that it must include a model of itself. If this is a good guess, why not try to design and implement more and more complex models of reality and install them volitrons?

Acknowledgement. This research was conducted as a part of the *Research on Human Communication* supported by the Telecommunications Advancement Organization of Japan.

References

Albus, J.S., Meystel, A.M. (2001) *Engineering of Mind: An Introduction to the Science of Intelligent Systems*, New York: J. Wiley & Sons.

Brooks, R.A. (1991) Intelligence Without Representation, *Artificial Intelligence Journal*, *47*, 139-160.

Brooks, R.A., Breazeal (Ferrel) C., Irie, R., Kemp, C.C., Marianovi_, M., Scassellati, B., Williamson, M.M. (1998) Alternative Esences of Intelligence, *Proceedings, 15th National Conference on Artificial Intelligence (AAAI-98), July 26-30, 1998, Madison, Wisconsin*, 961-968.

Buller, A. (2000) Self-Organization of Mind: Theoretical Background, Computer Simulation, and Empirical Grounds, Ph.D. Dissertation, Department of Psychology, Warsaw University.

Buller, A., Joachimczak, M., Bia_ow_s, J. (2002a) MemeStorms: NeuroMaze: A New Method of Pulsed Neural Network Synthesis, *The Seventh International Symposium on Artificial Life and Robotics (AROB 7th '02), January 16-18, 2002, Beppu, Japan*, 648-649.

Buller, A., Kaiser, L., Shimohara, K. (2002b) Meme Storms: A Cellular Automaton for Pattern Recognition and Dynamic Fuzzy Calculus, *The Seventh International Symposium on Artificial Life and Robotics (AROB 7th '02), January 16-18, 2002, Beppu, Japan*, 528-531.

Dawkins, R. (1976/1989) *The Selfish Gene*, Oxford: Oxford University Press.

Freud, S. (1920/1990) *Beyond the Pleasure Principle*, New York: W.W. Norton & Company.

Freud, S. (1923/1990) *The Ego and the Id*, New York: W.W. Norton & Company.

Nowak, A., Vallacher R.A.(1998) *Dynamical Social Psychology*, New York: Guilford Press.

Tulving E (1995) Organization of Memory: Quo Vadis? In: Gazzaniga, M.S. (Ed.) *The Cognitive Neurosciences*, Cambridge MA: The MIT Press, 839-847.