An Artificial Volition Architecture for Autonomous Robotics

T. E. Raptis[1] and K. Karamanos[2]

[1]*Division of Applied Technologies, National Centre for Science and Research "Demokritos", Athens, 15341 Greece.*
[2]*Physics Department, University of Athens, Zografou, GR-15784*

Abstract
We introduce a new computational architecture capable of exhibiting an archetypal type of volition in a simplified modularized version of M. Minskys hive-mind. In this model, three relatively independent computational cores which themselves can also be whole multi-agent systems are engaged in an endless interaction each one representing the internal "imaginative" world, the external world interface and the arbitrator or Internal Observer. Volition then is expected to occur as the result of an endless antagonism for control between the internal and the external world models.
Keywords A.I., Robotics, Volition, Free Will.

1 Introduction

History of A. I. is marked with a division between two main abstractions, the one of "connectionists" that try to imitate directly the real brain functions closer to the spirit of old Wieners Cybernetics paradigm and "symbolists" who advocate the old belief that mind is algorithmic, first introduced by Alan Turing. Both schools have had various successes in diverse fields but when it comes to the inner psychological or subjective experiences they face a major obstacle which is the correct objective definition and experimental verification of such abstract concepts as volition (exercise of will), consciousness and/or self-awareness.

These problems also arose in the early phase of development of Cognitive sciences and still constitute a major set of debatable issues at the heart of this research. Questions concerning the possibility of externally verifying ones own awareness of a Self and a personal identity have given rise to the argument of "Philosophical Zombies" against functionalist interpretations of the identity and awareness problem.Although the particular architecture proposed seems to be unique, at least to our knowledge, there are several parts that may be comprised into the proposed structure and have already been proposed as separate items and implemented into various robotic platforms.

A crucial property for self-aware agents is that of self-observation which requires a constant scan of both the history of the external agent behavior and the internal loop activity and their correlation. Recent work from the Cornell group[1-3] has revealed the possibility of an advanced reverse engineering of

complex environmental signals and accurate modeling of external reality. Such models can then be contrasted with the internal personality model and affect future planning. Furthermore, the agent can effectively randomize its own behavior to avoid stereotypical pre-programmed behavior by an internal mechanism which optimizes the antagonistic need to preserve the personality model and the environmental pressures by an adaptation protocol which affects the probabilities/weights of certain actions.

Attempts to introduce a "Self-model" compatible with the idea of an "Inner World" have been superficially carried out based on a 3 -dimensional Mental Space together with the so called "Big Five Model" of psychology. Previous work by the Waseda group[4-6] on a 3-D Mental Space has shown promising results in control of a robotic face (emotional agent). This way we can build "idiosyncratic agents" with a certain predefined set of preferences.

A second important step in the construction of intelligent agents is the previous work of Luc Steels on "Fluid Construction Grammars" which led to the development of the TALKINGHEADS project. In this, a number of bots were sustained inside the internet moving in a network of host computers which were equipped with steerable cameras. They were then able to analyze optical signals and classify objects while developing a primitive linguistic structure through which they were communicating their experience into one another. The same model can also be utilized in an advanced volitional agent who would become capable of classifying both internal and external signals and thus reaching at level 3.1 or 3.3 of Tables 1(a) - (b) of the next section. It could also be used for an agent applying an artificial linguistic structure to purely internal variables.

The present proposal attempts to provide an example of an advanced architecture that could encompass all previous developments and unify their separate approaches in a unique frame thus extending their capabilities towards an advanced agent design with truly inherent volitional attitude that would arise as a result of the internal dynamics of its major components. The presentation is based on a high level description of the whole architecture ignoring the details of the software implementation which in principle could be numerous. It is intended in giving a correct understanding of the foundations of an alternative volitional theory that could possibly be applicable in other fields as cognitive sciences and human psychology.

In section 2, we lay the foundations of our model while in section 3 we describe their possible implementation in more detail using a top-down approach. In section 4 we conclude and also discuss the significance of learning as an additional concept of a higher level that was not absolutely essential in the previous development.

2 Foundations of Artificial Volition

Adopting a practical approach, we choose to concentrate in the necessary and sufficient conditions for a behavioral evaluation of the existence of volitional attributes of an agent. Present state-of-the -art in cognitive sciences allows one to write a generic test that can be applied to an arbitrary set of agents in order to discriminate between several levels of awareness as proposed in [1]. This is summarized in the Table 1(a) below.

Table 1 (a) The awareness level

Awareness Level	Discriminating Question	Categorization
0	Is It animate?	Alive/Dead
0.1	"Does the system move or act on its own, i.e., without obvious prompting by external forces?"	Autonomous/Non-autonomous
0.2	"Is the systems spontaneous behavior modified by events/conditions in the environment?"	
1	"Does the system appear to be trying to approach or avoid any object or occurrence of an event in its environment"	
1.1	"Does the system have different sets of goals active during different environmental or bodily conditions?"	modal value-driven automaton/"Pac-Man Ghost"
2	"Does the system develop new adaptive approach or avoidance patterns over time?"	
2.1	"Can the system engage in a task that requires working memory (e.g. delayed non-match-to-sample)?"	
2.2	"Can the system engage in a task that requires long-term memory?"	
2.3	"Can the system engage in a behavior(e.g. game-playing, navigation) that requires evaluation of multiple possibilities without action?"	Most animals
3	"Does the organism send and selectively respond to social cues? "	
3.1	"Can the agent pick up and move around objects in its environment?"	
3.2	"Does the system communicate using language that has syntax as well as semantics?"	Chimpanzee

Some observations are due at this point. Nowadays, robotic arms and even autonomous robots exist that are capable of fulfilling 3.1 although they are no more than modal automata thus we should move this question into level 1. Secondly, the overall description of level 3 seems incomplete in that it does not explicitly shows true volitional acts in the absence of competition with a truly intelligent environment which includes also other intelligent agents either artificial or human. We thus propose the following modified table 1(b) for level 3.0.

Table 2 (b) The modified awareness level

3	"Does the organism send and selectively respond to social cues?"	
3.1	"Does the agent attempts to escape captivity enforced by an external agent or situation?"	
3.2	"Is the agent capable of enforcing a certain task to other less intelligent agents?"	
3.3	"Does the system communicate using language that has syntax as well as semantics?"	

We can now proceed to a separate examination of the three different levels. In fact we propose that the above hierarchy is not strictly necessary and that certain properties may be intermixed at least in artificial agents. That is to say that we can in principle separate between different architectures or implementations of systems that could emulate several of the characteristics pertaining at different levels without strictly obeying the above hierarchy. In particular, we would like to make a crucial separation between consciousness as a purely subjective state and volition as a more primordial level necessary for its existence. In fact, *it is not in principle possible to directly assert the existence of an internal "feeling" by the agent.It is only possible to assert the resistive actions taken by the agent against external obstacles or environmental "laws" that are not preprogrammed.*

Next we concentrate on a mechanism capable of exhibiting volitional effects at the level of 3.1 without necessarily reproducing all of the characteristics of the previous levels. Specifically, we seek for the minimal behavioral test that could verify the ability of an agent to exhibit an element of "free will" in the sense of a) either randomizing its own behavior in order to cope with a contradiction or a conflict between its own Self model and its obtained World model, or b) undertaking evasive actions against an enforced captivity or restrain by another agent. We thus concentrate on the fact that the presence of any kind of "free will" is definitely asserted only in an antagonistic situation where the agents will is exerted against another agent as a resistive force. This assumption does not exclude a cooperative behavior which would only occur under a state of agreement or symbiosis between different agents.

To explain the significance of the levels 3.1 and 3.2 we may also add a quote from a dialogue between C. Zvosil and D. Greenberger[8] where they criticize the incomplete nature of Turings test with respect to free will. *"· · · assume an artificial world and/or an artificial creature. Assume further a super selection rule which should not be broken by any circumstances; eg eating from the tree of Genesis. This creature should be termed intelligent if it breaks its super selection rule. In this approach it is evident that uncontrollability is a trade-off for intelligence and that it might be impossible to create a machine which is both intelligent and a*

reliable server." We choose to call this principle, a "Non Serviam" principle from the Latin expression for Denial of Service. We will next show that it is possible to extract from the above general argumentation a full computing architecture that can fulfill the above principle. The significance of such a construct is that a) if it is possible than it will probably be realized in the future yielding new problems for security and reliability of services (Friendly AI problem) and b) it seems to be more closely associated with human behavior than other mechanistic approaches.

At first we will use a metaphorical example to lay down the description of the principles of the model. The basic obstacle in present state-of-the art is that the mind is still represented much like an operating system which waits for input from the external environment and then reacts to it according to a prescribed set of functions. What is clearly missing from all these approaches is the notion of a purely internal world with autonomous life independent of any external activity. Such is the case of imagination but also of dreaming. The only way we can achieve a machine with a "dreaming state" is to endow such a construct with an internal autonomous and endless loop of which the variables are equally sensed by an "Internal Observer" (IO) as well as the external response signals albeit not always on an equal footing. We thus present this general idea with the following schematic.

In Fig.1, the three modules should be interpreted as relatively independent computational engines of different characteristics and purpose. While the Inner World (IW) is represented by an autonomous activity, like a huge dynamical system, the IO module is more like an operating system which needs to be fed with inputs from both subsystems having the environmental variables fed by sensory systems and also form the interface of the machine with the outside world. Each module may act as an independent agent in a continuous dialogue with the other two thus forming a tri-dialogical system.

The dynamics envisioned behind this scheme can be described with the aid of Fig.2. The IO module acts like an independent agent "trapped" between the activities of the IW core which is fed to it through appropriate internal sensory inputs and those provided by the external interface. The IO module then attempts an endless evaluation and arbitration task between the conflicting activities in order to control the appropriate responses that are also imposed by the third Survival Task (ST) module which plays the role of an autonomic nervous system. In what follows we give precise meaning and some possible technical methods to implement the above logic. We of course assume the existence of an appropriate environmental interface for both input and response signals (eg. a full robotic body).

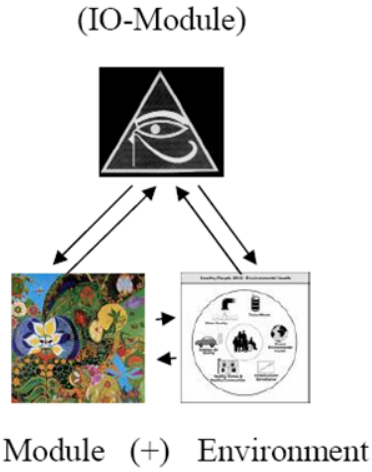

Fig. 1 A tri-dialogical system

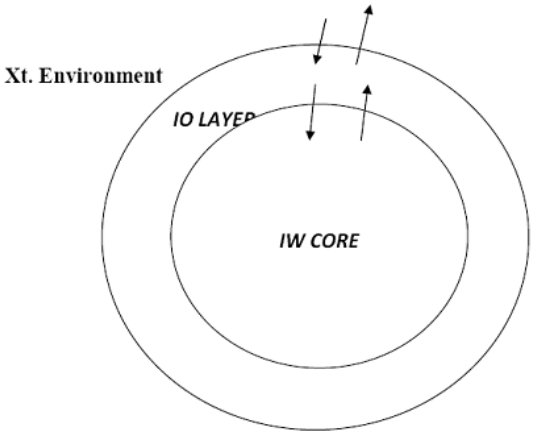

Fig. 2 A dynamic response process

1. The IW-module is a specially constructed closed computational loop which is to have access to some or all the external inputs not for processing but in the form of direct perturbations of its own internal variables. It is also prewired in a certain way with the IO module in a way that allows for alterations of its internal structure -not just its variables!-imposed by the IO module. (Practically this means that the IO module may alter even the form of the "equations of motion" of the internal dynamics). This is supposedly an endless internal activity. At

this point one may ask how we can guarantee that this will indeed be an endless computation without solving the halting problem! This will be further clarified into the next paragraph that fully describes the dynamics with the environment mediated by the IO activity. At the moment it is important to accept the IW module as an isolated "dreaming" machine or a kind of "subconscious" in the overall architecture.

2. The ST-Interface module contains all such attributes that represent predetermined protocols for the survival of the agent in an arbitrary environment including input signal processing and automatic response circuitry. It is thus close to a model of the autonomic nervous system.

3. The IO is by necessity the part of the machine responsible for the categorization and evaluation of the activities of both the internal and external environmental variables. It thus holds and updates a primitive Self-model made by the subsequent observations of each ones activity. On the other hand, it also plays the role of an arbitrator between possible conflicting demands of the two separate IW and ST agents in case of conflicting demands posed to the external response modules.

In order for this scheme to exhibit full volitional attributes there is one more crucial step in the definition of the IO module. We thus propose to introduce an "egotistic" attitude to the agent in the following sense. The fundamental property of "Ego" is not to be just a Self-model in the form of a Data Structure but also to be "demanding". This is also evident from early childhood psychology. The only way we can give a precise technical meaning to this proposition is to introduce a certain kind of internal Hyper-Tasks attributed to the IO module beyond the simple self-identification task. We thus attempt to derive such a Hyper-Task from a Generalized Functional Control problem. We assume that the simplest Hyper-Task is the attempt by the machine to create a "higher Self" model into which all or as many as possible of the external variables have been "internalized" or in fact "enslaved". This is to be understood here in a somewhat more broad manner than a "master-slave" configuration is often realized in engineering. This is quite a broad concept that can extend even to systems with continuous variables. We interpret this internalization process as follows.

Assume a composite system, e.g. a Neural Network, which has several processing nodes and several other input nodes (Fig.3). The processing nodes are to be considered as "immediately active" in the sense that they carry a certain processing capacity while the sensory nodes form an interface for some in principle unknown external nodes that represent the activity of the environment. Thus all nodes f1, f2, $\cdots$, fN represented by circles are "known" to the system while the nodes e1, e2, $\cdots$, eM represented by red squares may not even have a functional expression due to their extreme complexity. In a sense, although the agent may

possess a primitive notion of identity it still considers the rest of its world as an extension that can be manipulated. The systems dynamics executes the Hyper-Task of trying to minimize a generalized distance between the functional forms of the external world source nodes and the internal nodes. Thus the system should start to transform the external world continuously until it would be able to bring all or most of the external variables to a state in which they conform to the internal world dynamics. While this in actual reality is an open and endless task (considering the rest of the universe as the "environment"!) it is still an accurate representation of the origin of the agents volition. In fact, we propose that this is the only mathematically acceptable expression of an agents "free will". An interesting consequence of the above type of dynamics that we can predict is that cooperation with such an agent would be possible through "enticement" which presupposes a degree of access or knowledge to its own internal variables that make up its "character" or Self-model without sacrificing an amount of fluidity in it. Conclusively, we may end this section with the following propositions.

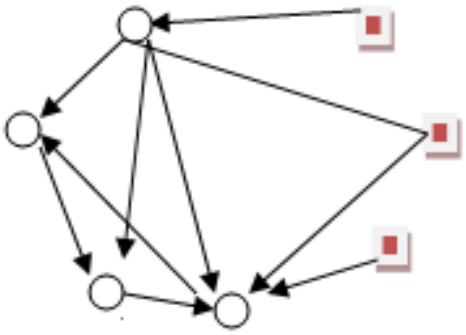

Fig. 3 The internalization process of artificial volitional attributes

Proposition 1. *A sufficient condition for the appearance of volitional attributes in a network of parallel Turing Machines is the existence of conflicting dynamics between computational tasks of different purpose that result in the formation of a master-slave configuration inside the network.*

Proposition 2. *A necessary condition for ascertaining the volitional attribute of an agent is the testability of its capacity to interpret restrictions as such even in the absence of any knowledge about their origin ("suspicious agent") and seek more freedom in the transformation of a given environment.*

3 Possible implementation and dynamics

A more concise and detailed description of the system units, their interconnection and dynamics follows. The design philosophy explained in the previous section

is further elaborated on the technical details of a possible implementation.

Design of the "Inner World" model can be better described with the aid of the analog computing paradigm which is here simplified on a system of differential equations. These can be seen as an average over a multitude of microscopic degrees of freedom that could be realized by neuromorphic circuitry. We believe this to be true for natural biological circuits although we do not make strict use of neural implementations as this is unimportant to our purpose.

Let then x1,$\cdots$,xnin and x1,$\cdots$,xnout be a set of internal and external variables defined as follows : the internal variables form the basis of the inner loop dynamics while they are coupled to the set of external variables. The external variables are coupled to environmental signals and they originate at the ST module where a first processing of the system inputs takes place. System response is due to both the internal dynamics of the ST module as well as the supervising inputs from the IO module which mediates between the IW module and the ST module.

The inner loop attempts to bring the set of external variables under its own control. This of course is an always incomplete task in the sense that the external variables obey an in principle unknown, non-stationary dynamics which may also have different underlying laws from a previous instant to the next. The philosophy behind this control scenario is that *the inner loop attempts not just to identify the external variables dynamics but it tries to "enslave" them in order to follow its own internal differential equations. That is to say, the inner loop attempts to internalize the external world.*

It is possible to borrow from nature the additional well known biological principle of antagonistic signals. For this, we would have to define pairs of antagonistic variables. One can utilize for example the well known model of Competitive Lotka-Volterra equations to simulate the inner loop dynamics. The choice of the system is somewhat arbitrary and has been chosen for having stable attractors. One could in principle try other choices or a direct neural implementation. Even a spiking network could have been used but at the cost of an increased complexity.

In fact, we may assume the continuous version of a PDE system or equivalently the corresponding pseudo-spectral kernel by which the IW module builds the subsequent configurations of its own internal state. Assuming an appropriate sampler interface between the IO and the IW modules, the IO "sees" snapshots of the IW activity the same way it sees the external world through the additional interface of the ST module. Moreover, the inner loop module should be coupled to a set of supervising signals from the IO module that affect a matrix of coefficients for the ODE/PDE system that defines the IW loop dynamics.

The inner loop dynamics may be made to obey a 1st order Hyper-Task. This

is described by an internal update process of the system coefficient matrix or of the corresponding kernel parameters towards the production of more coherent and symmetric patterns of activity (artistic agent). This is also influenced by the perturbations from the external variables so that an antagonistic dynamics between the inner and the outer world models is established. The criteria of what constitutes a like pattern are prewired into the particular agent and form an integral part of its "character". (Martian spiders for example may have a very peculiar idea of what constitutes a beautiful face!)

The IO module is primarily responsible for a 2nd order Hyper-Task of reducing the functional distance between the IW and the Outer world as it appears to the interface variables provided by the ST module. This 2nd order Hyper-Task is eventually linked to the 1st order one through the attempt to bring the environmental dynamics close to fulfilling the directive of the IW module of producing more symmetric and coherent patterns. In principle there is no limit in the hierarchy of Hyper-tasks that could be implemented and could also be said to stand for the agents "talents" in a metaphorical sense.

The IO module should contain a reverse engineering engine which separately examines time-series of both the internal and the external variables and their history which are recorded into a dedicated part of memory (long term memory). It should then extract mathematical models of both and attempt a possible match by proposing certain changes into both the internal parameters and the external assumed environmental parameters. The supervising signals towards the IW module can then be directly implemented into the underlying system of differential equations while the supervising signals towards the ST module must be appropriately translated into actions through an interpreter that extracts from them the discrete signals towards the drivers of the external devices in a manner that can best allow the modification of the assumed external dynamical functions.

Such a modification of the external world may also obey certain directives like the overall entropy reduction in the surrounding space. Other scenarios can also be tested with more complicated directives. In fact, the IO module can be updated to include a more general planner under certain learning strategies. For the moment we may ignore this advanced capability as it is not inside our objectives to test learning strategies of which are many in the trade.

The IO module is in addition responsible for continuously building and refining a Self model by projecting the present state of match or distance between the IW and the external world model into the Mental Space representation which may affect the decision making and subsequent planning. In total, the IO module is responsible for all the abstract representations of both the Self and the external World. The IO module can also incorporate a FCG engine for artificial linguistics in order to fulfill the final level of self-expression.

The ST module must contain certain instructions for preprogrammed tasks like step generators for locomotion, or other necessary movements, auto-charging in case of a power source present, "food" hunting if abscent etc. In general it may be characterized as the equivalent of the autonomic nervous system. On the other hand, the ST module must interpret the supervising signals from the IO module in order to feed appropriately the drivers of the external peripheral devices (legs, arms, sensors) so that the combined motion will affect the environment in a way fit to the higher task of enforcing the desired structure to the external world dynamics.

Building the appropriate interpreter for the drivers is a non-trivial task as it requires also an amount of computational geometry in order to take into account the exact actual details of the environment and the robotic devices and the possible ways of their interaction. At this level, it is necessary to include a system of somatosensory perception which entails the robot with the capability of describing itself in both shape and movement inside a certain environment.

4 Discussion and Conclusions

The above general scheme is here presented as having certain advantages over other existing approaches due to its holistic nature that attempts to incorporate the most fundamental elements of what human beings intuitively know about themselves. To our opinion these advantages include.

- Deeper understanding of the foundational principles behind awareness in artificial and natural systems.
- Enhanced capabilities of operation of autonomous systems in real time.
- Practical approach to the problem of measuring various forms of awareness in situ.
- Prediction of possible malignant applications of true self-aware software or hardware and experimentation with counter-measures ("Asimov Laws" - ONR Report[9]).
- Practical investigation of the problem of Friendly AI (HAL9000 problem).

In principle there seems to be no restriction in the kind of Hyper-Tasks that could be implemented. What has not been examined in detail here due to space restrictions is the role of learning processes in a direct interaction with the IO layer. It seems reasonable to assume that a learning machine with the structure described above could also discover its own set of Hyper-tasks or modify previously programmed ones unless this is strictly forbidden via some dedicated censorship circuitry (Freudian viewpoint).

With respect to the last two points above it deserves to mention that there

exists indeed a possibility that a learning machine could also have been imprinted - or even develop on its own! - a "seek and destroy" attitude. In fact, it is quite possible that a "Skynet" scenario is in principle technically feasible and could become reality in the next 20 years or so as already predicted by De Garis[10] and Kurzweil[11]. The architecture presented above to our opinion justifies such fears and shows the necessity towards more research in the direction of friendly AI as well as the need to ask for demilitarization of robotics.

Acknowledgements:

The first author would like to express his gratitude to the members of the Computational Applications Group of D.A.T.-NCSRD for helpful discussions and support in the preparation of this document.

References

[1] Bongard J. and Lipson H. (2005) "Active coevolutionary learning of deterministic finite automata", *Journal of Machine Learning Research*, 6(10):1651-1678.

[2] Bongard J., Zykov V. and Lipson H. (2006), "Resilient machines through continuous self-modeling", *Science*, 314(5802): 1118-1121.

[3] Bongard J. and Lipson H. (2007), "Automated reverse engineering of nonlinear dynamical systems", *Proceedings of the National Academy of Science*, Vol. 104, No. 24, pp. 9943-9948.

[4] Miwa, H., Umetsu T., Takanishi A. and Takanobu H. (2001), "Robot personality based on the equations of emotion defined in the 3D mental space", *Proceedings of the 2001 IEEE International Conference on Robotics and Automation*, pp.2602-2607.

[5] Miwa, H., Umetsu T., Takanishi A. and Takanobu H. (2001), "Human-like robot head that has personality based on equations of emotion", *Preprints of the Sixth Symposium on Theory of Machines and Mechanisms*,pp.1-8.

[6] Miwa, H., Umetsu T., Takanishi A. and Takanobu H. (2000), "Robot personalization based on the mental dynamics", *Proceedings of the 2000 IEEE/RSJ International Conference on Intelligent Robots and Systems*,pp.8-14.

[7] Chadderdon and G. L. (2008), "Assessing machine volition: an ordinal scale for rating artificial and natural systems", *Adaptive Behavior 16*, pp.246-263.

[8] Svozil, K. (1993), *Randomness and Undecidability in Physics*, World Scientific.

[9] Lin P., Bekey G. and Abney K. (2008), *Autonomous Military Robotics: Risk, Ethics, and Design*, Ethics & Emerging Technologies Group, California Polytechnic State University.

[10] De Garis H. (2005), *The Artilect War*, ETC Publication.

[11] Kurzweil R. (1990), *The Age of Intelligent Machines*, MIT Press.

Corresponding author
T. E. Raptis can be contacted at: rtheo@dat.demokritos.gr