ARTIFICIAL VOLITION:
IS IT EVEN POSSIBLE?

Theophanes E. Raptis
Computational Applications Group
Division of Applied Technologies
NCSR Demokritos 2015

What is Volition/"Free Will"?

- Is/Can there be an effective algorithmic definition?
- Is it independent from Social Relation (Not individually definable)?
- Useful Abstraction: Distinguishment between
  - Anthropomorphic (Evolutionary, Self-sustainment oriented, Culturally motivated)
  - "Machine" Intelligence (Manifestation of an emergent order from preprogrammed elements)
- Both expected to be distinctly purposeful (plan/task oriented, "meaningful") instead of mere random choices from some 'learnable' probability distribution.

HOW TO (MEASURE/REPRODUCE)

- We do not fully understand/"grasp" a truth before being able to 'reproduce' what we learned.
- Would you trust an engineer that claims knowledge of bridge building before having built any single one?
- Would you not let engineers make 1st time rockets for the moon?!
- Evolutionary Trade-Off: Risk vs Gain
- Is Volition Measurable?
- Is Volition Constructible?
- What-the-blip do we understand of our own 'Will'?

General Awareness Criteria (and Their Volitional Complement)

- ***Physical Presumptions:***
- Classical/Macroscopic (Strong Decoherence Regime – Is it obligatory?)
- Deviation from Variational-Least Action Principles (Non-Mechanistic/beyond mere Control Theory)
- ***Presently Established Cognitive Criteria:***
- Classification Based on a set of fundamental questions related to phenomenological observation
- Animate/Inanimate.
- Autonomous/Nonautonomous.
- Internal/External Causation.
- Philosophical Zombie Problem/"Pack-Man Ghosts"/Modal Value-Driven Automata.

[Chadderdon, G. L. (2008). Assessing machine volition: an ordinal scale for rating artificial and natural systems. *Adaptive Behavior*, 16(4), 246-263]

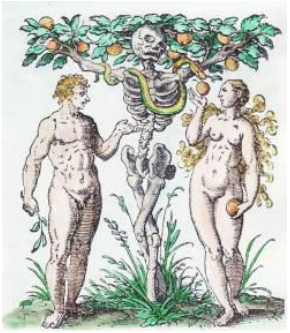
- ***Hierarchy of Automata*:**

- Level 0: Non-Volitional Systems
- Level 0.0: Inanimate Objects
- Level 0.1: Schizoid Automata
- Level 0.2: Reactive Automata
- Level 1: Instinct-Driven Automata
- Level 1.0: Value-Driven Automata
- Level 1.1: Modal Value-Driven Automata
- Level 2: Contained Self Organisms
- Level 2.0: Pavlovian Organisms
- Level 2.1: Ideational Organisms
- Level 2.2: Recollective Organisms
- Level 2.3: Deliberative Organisms
- Level 3: Extended Self Organisms
- Level 3.0: Social Organisms
- Level 3.1: Manipulative Organisms
- Level 3.2: Symbolic Organisms
- Level 3.3: Cultural Organisms

Critique of the Artificial/Natural Dichotomy

- Poor physicist's rationalization: Artificiality as an Anthropomorphic Category.
- Remnant of religious (Godly) "Prototype" vs Material "Copy".
- Need for a new logical category based on "Constructibility".
- Complementary to "Computability".
- Humans/Machines as members of the class of (evolutionary) constructible entities.
- Recently proposed "General Theory of Constructors". (D. Deutsch, C. Marletto, S. Benjamin, http://constructortheory.org/)

THE *"NON-SERVIAM"* PRINCIPLE



- Main Motive behind this work

"Assume an artificial world and/or an artificial creature. Assume further a *"Superselection Rule"* which should not be broken by any circumstances; e.g. eating from the Tree of Genesis. This creature should be termed "intelligent" if it breaks the superselection rule. In this approach it is evident that controllability is a trade-off for intelligence and that it might be impossible to create a machine which is both intelligent and a reliable server." *D. Greenberger* quote to *K. Svozil*, *"Randomness and Undecidability in Physics"*, World Sci. 1993.

- In simple terms, "truly Free Agents contain a *rebellious* attitude".
- In empirical terms, "would you not consider a child that never breaks any rule, incapacitated?"
- Is true discovery possible without a degree of insubordination?

Non-anthropomorphic General Volition

- Free Agent: Self-Imposed causation (true deviation from mechanistic behavior.)
- Freedom of *prescribed* choice not enough: creation of new "choices" – "*Imaginarium Principle*".
- Internal causation unrelated to "laws of motion".
- Requirement for redirection of "Attention" to different internal thoughts instead of external environmental signals.
- Hence: *Requirement for an Inner World*.
- Machines with a *"Dreaming State"*.
- *Cybernetic Systems that can keep working even in the absence of any "environment" thus working on the contents of "themselves"!*

The Art of Inner World Building

- The *"Solaris"* parable.
- What is the opposite of a standard Operating System?
- Self-surveyed closed loops.
- Computations with & w/out an *"All-Seen-Eye"*.
- Local vs Non-local operations → gradient vs integral operators acting on a matrix.
- Take any standard *Cellular Automaton* under an arbitrary set of rules.
- Couple the rule set with an integral operation like the average of all its states and you immediately end with…

Self-Observation - Self-Reference

- Self-Observing Closed Loops as Meta-Mathematical objects.
- The *"Strange Loop"* principle(D. Hofstadter, *"Goedel, Escher, Bach: An Eternal Golden Braid"*, 1979, *"I Am A Strange Loop"* 2007)
- Assume a large automaton of networked elements (extension of C.A., N.N. paradigms)
- Assume a special set of non-local, integral operators (NLO) acting on its set of states at every discrete time-instant t_i.
- Assume an additional "*evaluator*" circuit trying to isolate certain interesting emerging patterns using special measures based on the set of NLOs.

- Assume further, a large memory stack for previous images of memory maps storage.
- Assume also the capacity of the evaluator to change its own protocol based on some learning mechanism from previous experience.
- Last requirement leads to a type of meta-programming.
- Possibilities to be explored:
- Capacity of such a system to set up an internal multi-agents game in order to increase the overall experience contents of the evaluator.
- Extensions of the Solaris parable.
- *Hyper-Tasks* and Multi-valued forms reduction as meta-mathematical operations.

"Waking up" The Agent

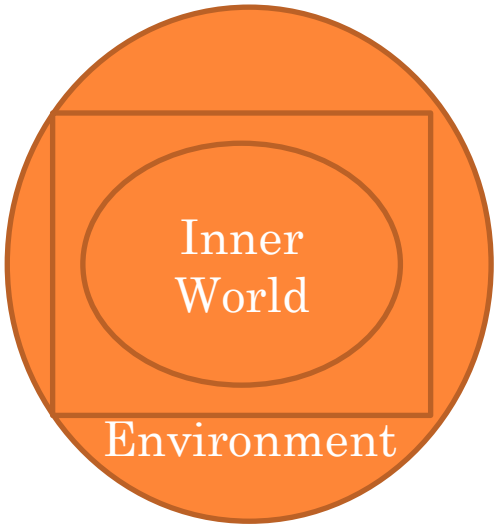
- Putting the external world into the game.
- When is a free agent rational?
- Rationality of complex agents always subjective without direct access to the true contents of internal world representation causing motivation.
- Assume a certain network-like representation of the agent's inner world model.
- External world representation reduction to a larger network of which the agent itself is a part.
- External nodes represent unknown functions.
- Hyper-task: Conquer all external nodes! (enlargement of 'Self'.)

"Conquistadores"

- Birth of "meaning" by formation of evaluation protocols during self-inspection.
- Valuation of internal imagery.
- Comparison with unknown nodes activity.
- Extension of internal meaning by attempt to subdue external nodes.
- Enlargement of internal valuation through learning of new external responses.
- Transformative Agents.
- Attempt to modify external environment in order to conform with internal valuations as a first "objective" indication of volition.
- Translation of agent's will uncertain (Quine).

Arbitration in Rational Agents

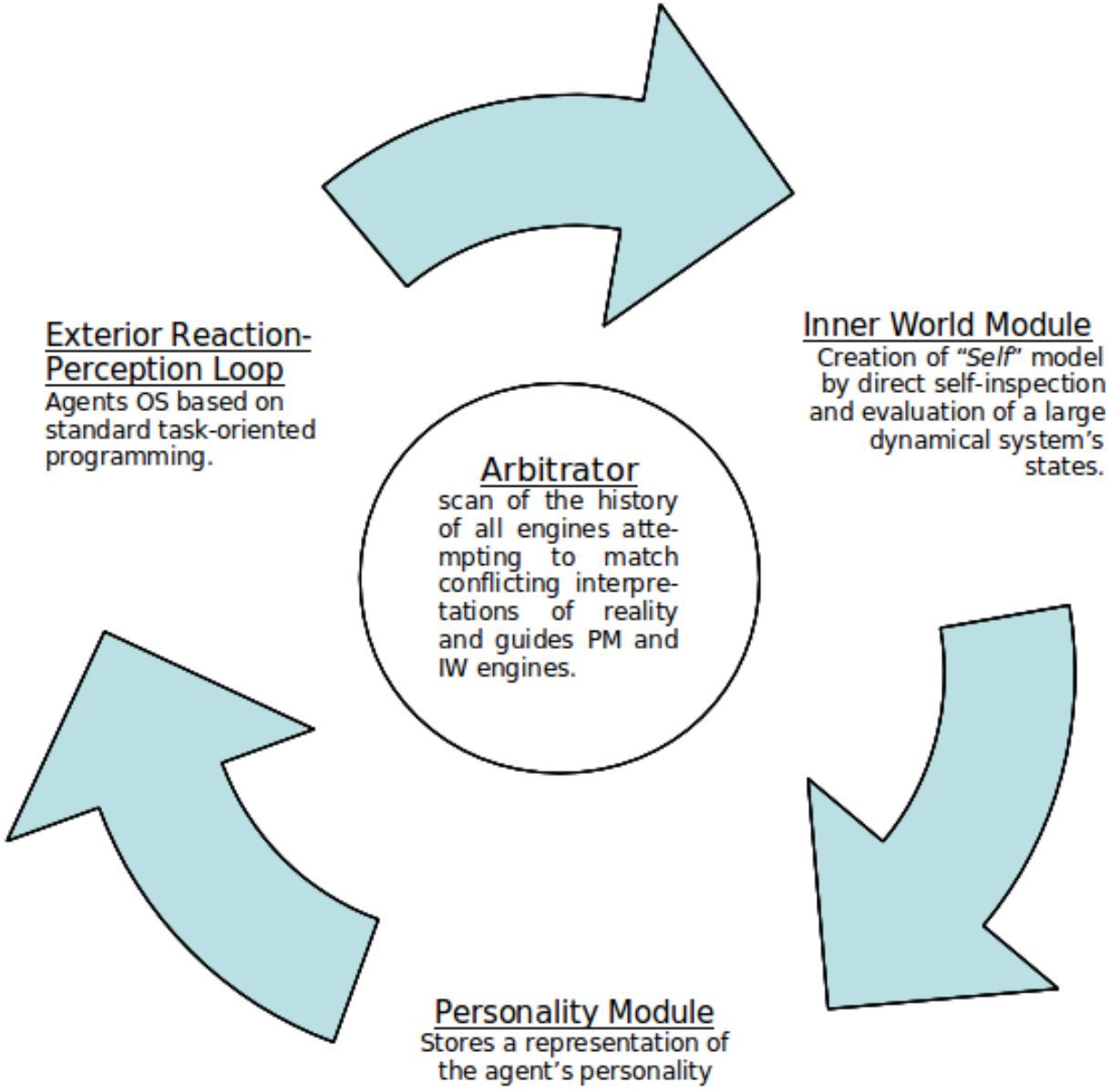
- Application of the evaluator scheme in a real environment requires a symbolic logic layer.
- In a human created agent, some capacities can be prewired/preprogrammed (vs evolutionary selection).
- Need for an Arbitration Interface.
- Choices weighted under physical and logical demands e.g., self-sustainment/survival.
- Understanding "Choice" through analysis of multi-valuedness and bifurcations.



Treatment of Bifurcations via Global maps

- Assume an internal symbolic alphabet $\{s_0,\ldots,s_n\}$.
- Assign a word $\{s_i\}^k$ to each set of actions causing environmental response.
- Hierarchical reduction protocol
- Given a multi-valued response associating each word with an ambiguous output, label each new path with a new symbol.
- Rebuild the combinatory powerset of already known maps with the enlarged alphabet $\{\{s_0,\ldots,s_n\},\{ l_{n+1},\ldots, l_{n+m} \}\}$.
- Rebuild tree of actions-reactions and repeat whenever necessary.
- Combination with probabilistic inference possible.

Schematic of a Possible Architecture



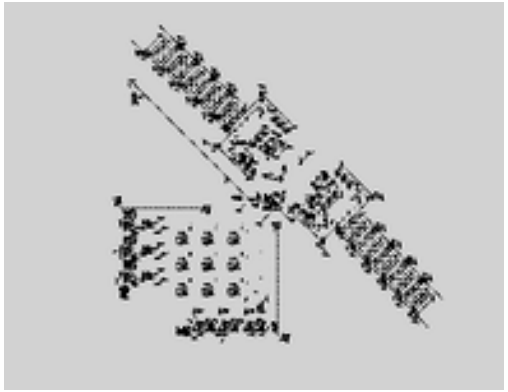
Volition as Meta-computation

- Restrictions of ordinary Turing Machine model.
- Made for prescribed computations on numbers.
- Not suitable for interacting computational engines exchanging messages that alter the evolution of computations.
- Main alternatives:

- Mind as an "Internet" (Minsky's *"Hive Mind").*
- *Interactive Computation Paradigm.* (D. Goldin. S. A. Smolka, P. Wegner, *"Interactive Computation*: *The New Paradigm*", Springer, 2006)
- Self-Rewriting Systems.
- Schmidhuber's *Gödel Machine*: Self-referential Machine Learning.
- Massalin's *Superoptimization*.

Imaginarium: HOWTO

- What type of dynamical system?
- Computational Universality: Neural Netwroks, Cellular Automata.
- A UTM simulated by Conway's "*Game of Life*".



- Extension and generalization of C. A. paradigm possible after careful analysis of the interaction type and the role of dimensionality in interacting discrete topologies.
- Combine with *Darwin Machine* paradigm.

- Currently tested: Networked *Collatz* elements as computational models of spiking neurons, Automata based on *Hasse* diagrams introduced by Rodriguez to simulate Spin-1/2 dynamics.
- Alternative one-dimensional representation of any discretised automaton with integer alphabet.
- Dimensionality transferred to an interaction kernel.
- Equivalent continuous dynamics

$$\partial_t c(x) = F\left(\int dx' K(x,x') c(x')\right)$$

- Interesting patterns associated with self-similar transfer functions/"rules" for F.
- Testing/training of F and kernel type during self-inspection by the agent.

Biological Analogues

- Ion Channel switching on cellular membranes with nearest neighbor micro-current coupling

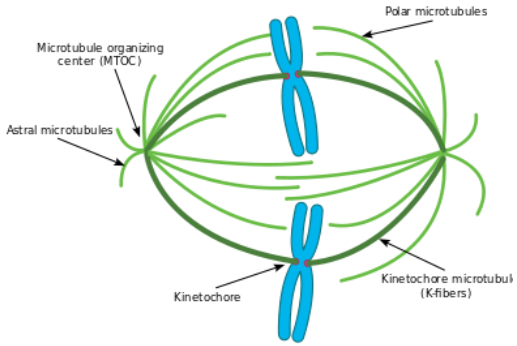
Patch-Clamp recordings

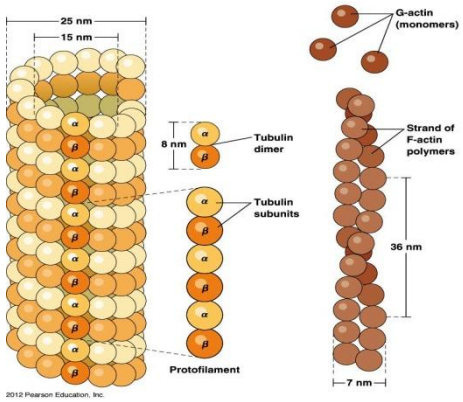
$$j_{\mathrm{A}} = -D_{\mathrm{A}}\left(\frac{d[\mathrm{A}]}{dz} - \frac{n_{\mathrm{A}}F}{RT}\frac{E_m}{L}[\mathrm{A}]\right)$$

Goldman equation

- Microtubular activity

$$E_m = \frac{RT}{F}\ln\left(\frac{\sum_i^N P_{M_i^+}[M_i^+]_{\mathrm{out}} + \sum_j^M P_{A_j^-}[A_j^-]_{\mathrm{in}}}{\sum_i^N P_{M_i^+}[M_i^+]_{\mathrm{in}} + \sum_j^M P_{A_j^-}[A_j^-]_{\mathrm{out}}}\right)$$





- Could these serve as an IC automaton evaluators based on Spintronics?

A Brave New World Ahead

Thank You!