Signs of consciousness in humans and machines

Włodzisław Duch

Uniwersytet Mikołaja Kopernika
Katedra Informatyki Stosowanej
Laboratorium Neurokognitywne ICNT

Google: W. Duch

Zjazd Filozofii Polskiej, Poznań, 18/09/2015

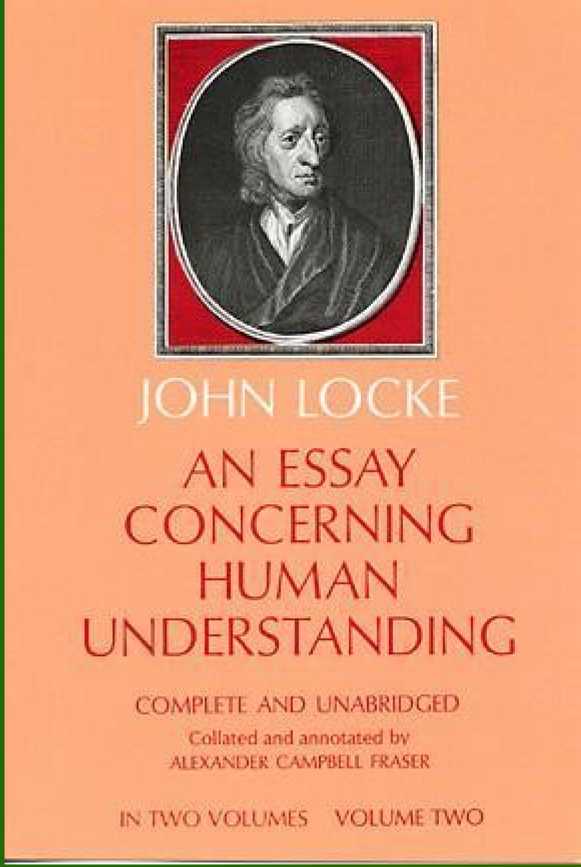
Consciousness

John Locke, *An Essay Concerning Human Understanding*, 1689. Book II, Chap. I, §19

Consciousness is the perception of what passes in a man’s own mind.

Questions:

1. What is perception? What am I conscious off?
2. Can the same mechanism be implemented in artificial systems?
3. What are the perspectives to build conscious machines? One can’t avoid neurophilosophy here.



JOHN LOCKE

AN ESSAY CONCERNING HUMAN UNDERSTANDING

COMPLETE AND UNABRIDGED

Collated and annotated by
ALEXANDER CAMPBELL FRASER

IN TWO VOLUMES VOLUME TWO

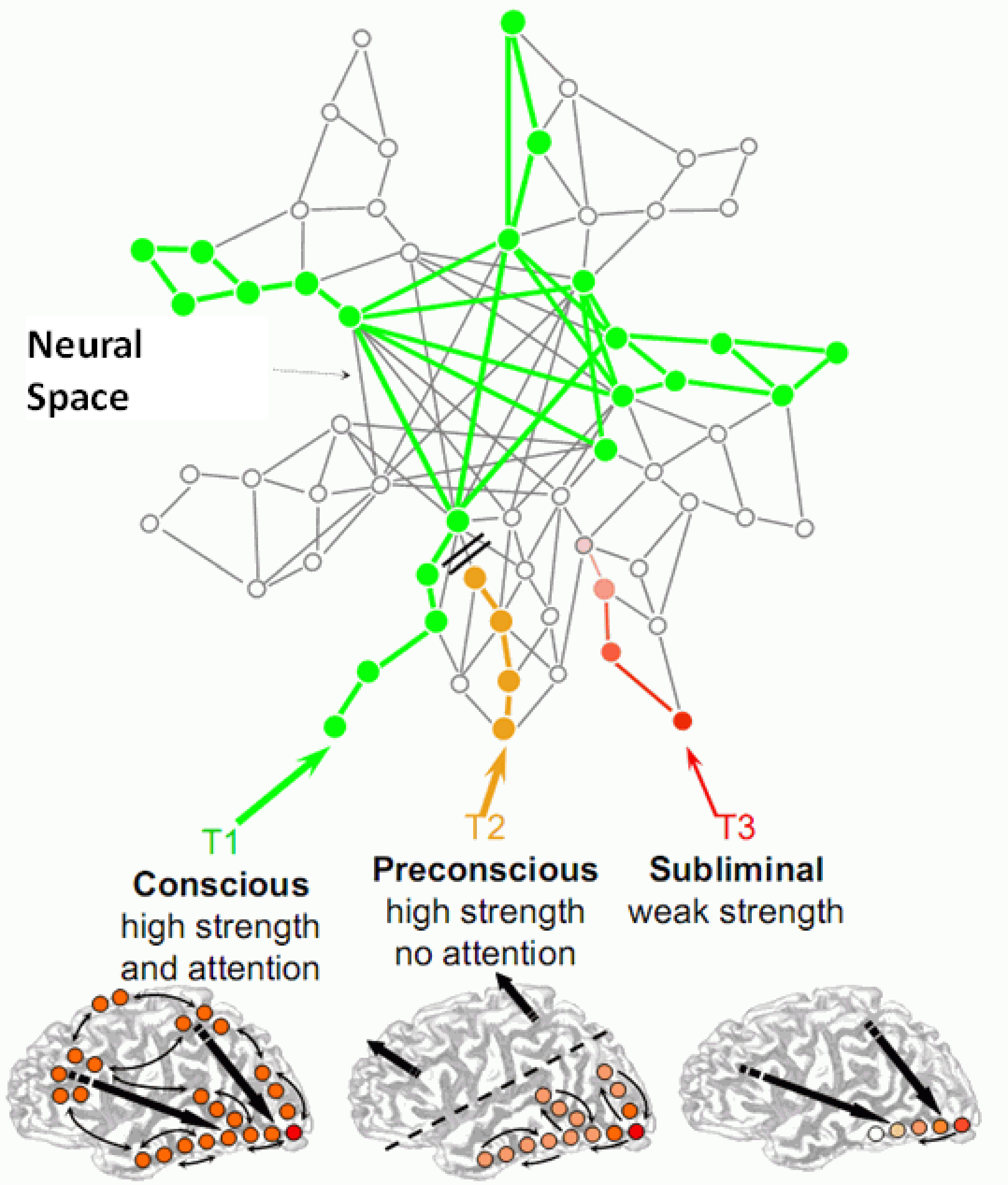
Conscious Perception

Very little of what passes
In the brain is perceived.

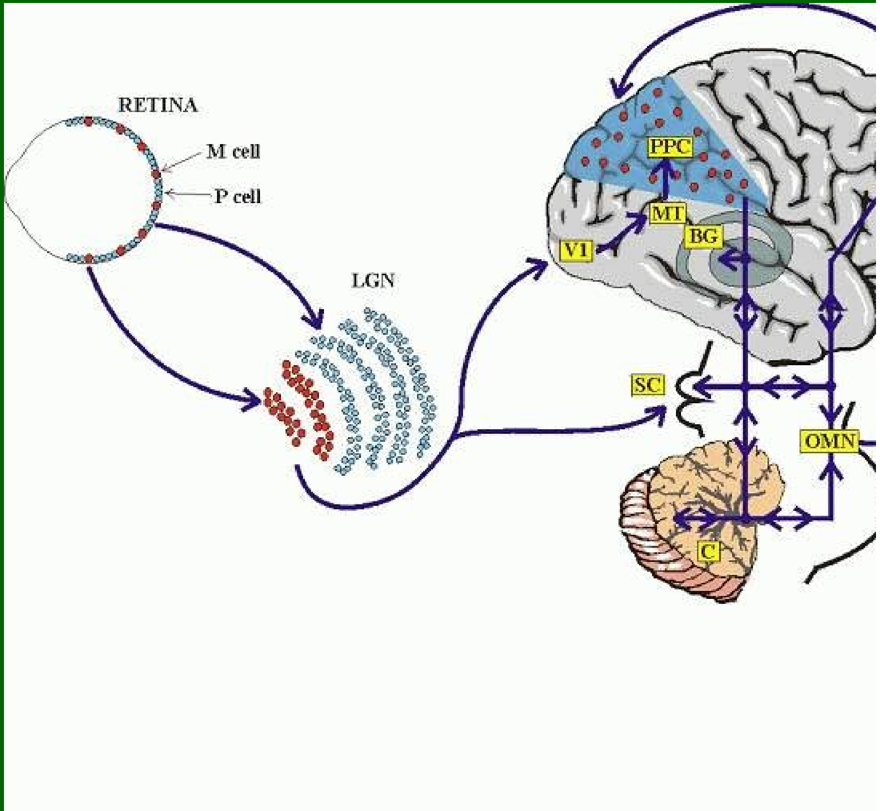
Attention + stimulation is needed to create brain states that are persistent and can be distinguished from noise.
Attention: 20 Hz
Perception: 40 Hz

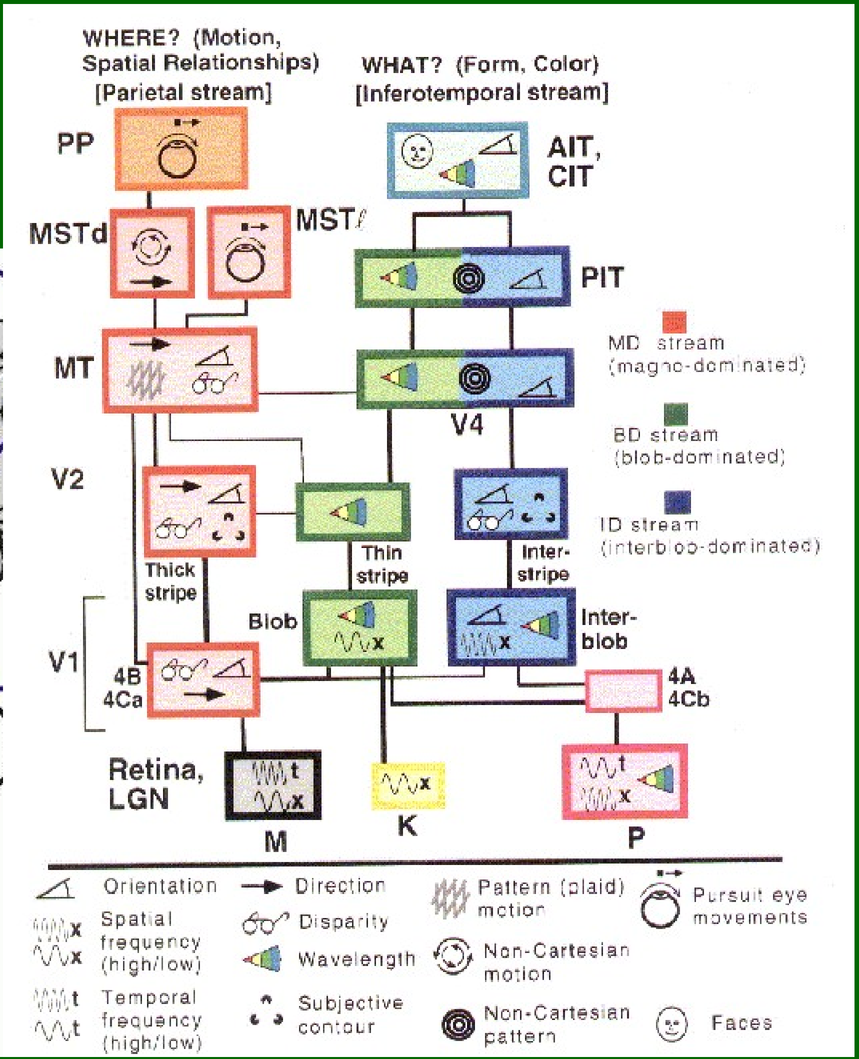
C. Gilbert, M. Sigman, Brain States: Top-Down Influences in Sensory Processing. Neuron 54(5), 677-696, 2007

Dehaene, Changeux, Naccache, Sackur, & Sergent, TICS, 2006

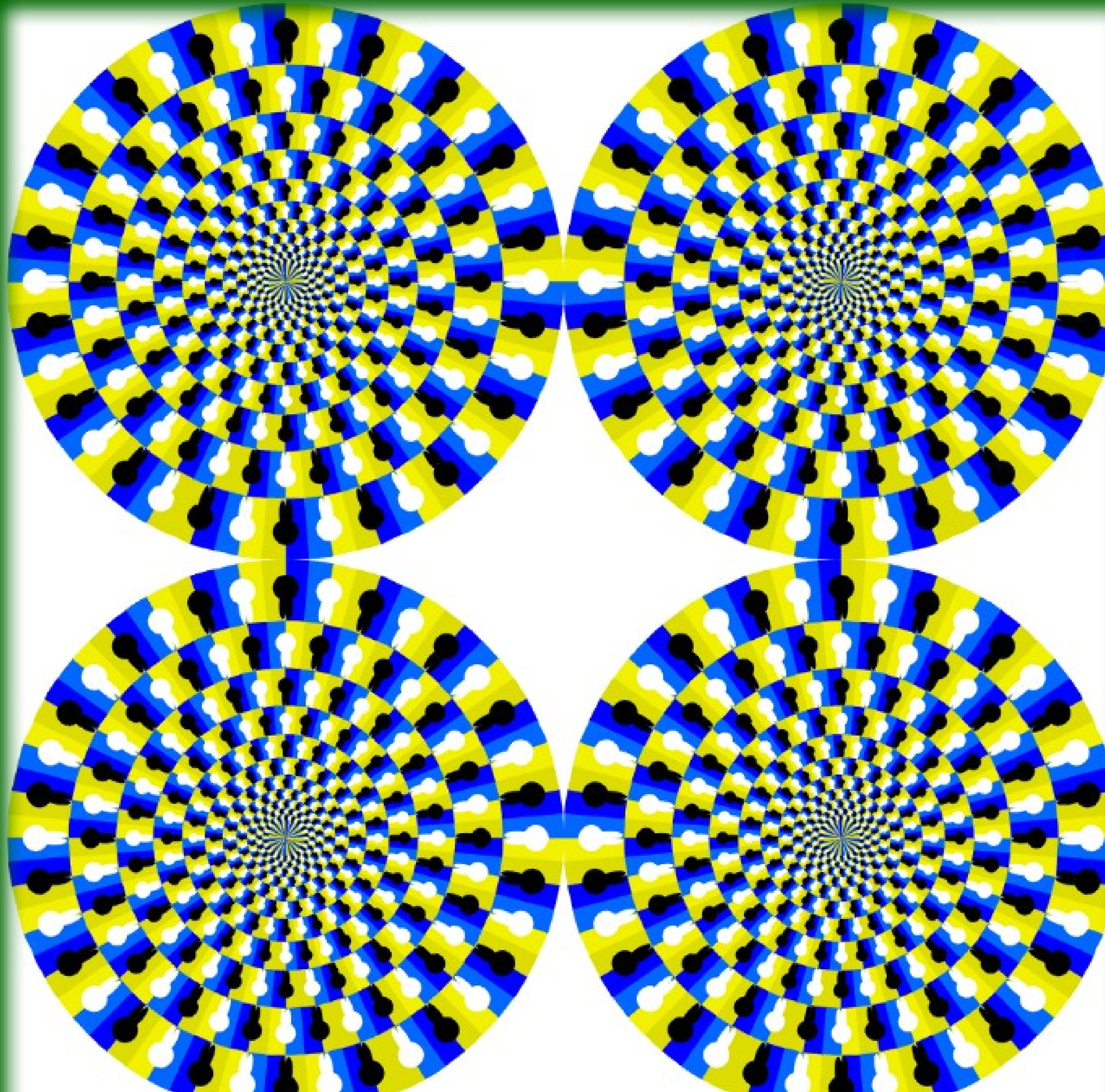


Vision

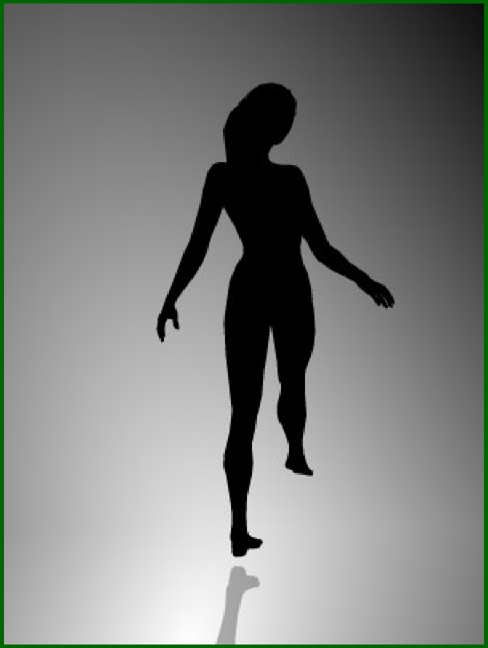


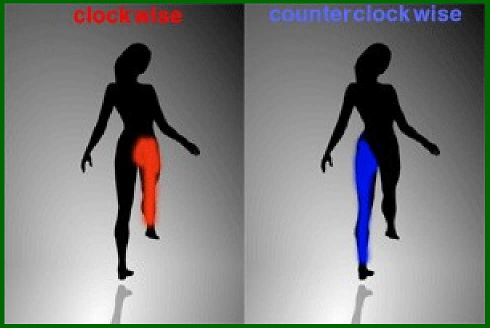


- How far does the signal from retina gets?
- If it creates strong, persistent state, stable for at least fraction of a second other parts of the brain may act on it, categorize it, initiate motor response, make a verbal comment, follow association.



It is your mind that moves



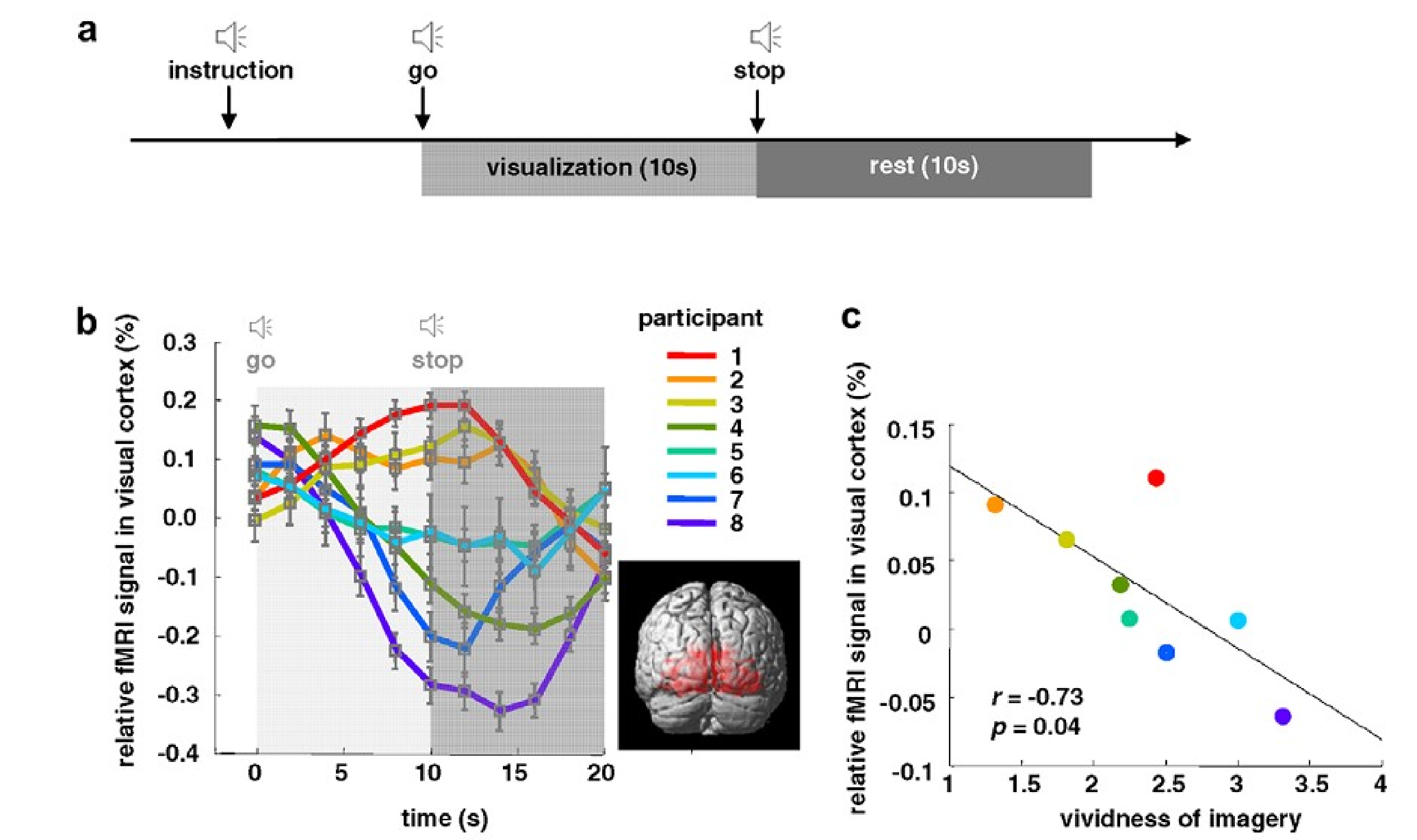


"Whilst part of what we perceive comes through our senses from the object before us, another part (and it may be the larger part) always comes out of our own mind."

William James, The Principles of Psychology, 1890

Attention to details

Content of conscious perception is in the whole brain.



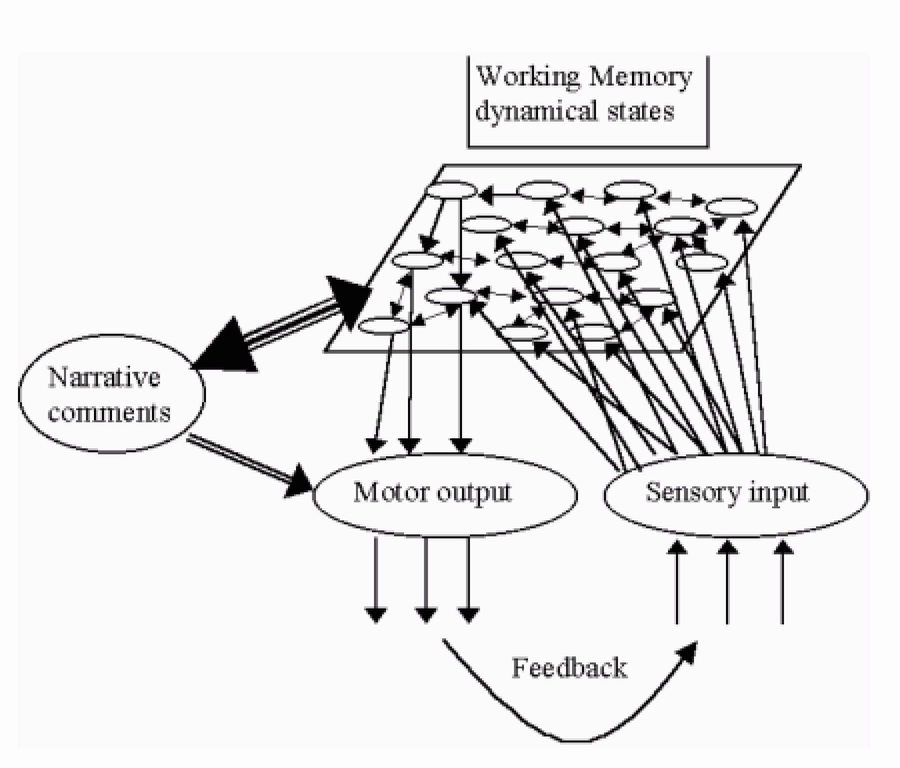
Results of the Vividness of Visual Imagination (VVIQ) questionnaires and V1 activity measured by fMRI are strongly correlated: some details are in V1.
Cui, X et al. Vision Research, 47, 474-478, 2007

Brain-like computing

Understanding requires simple mental models.
Brain states are physical, spatio-temporal states of neural tissue.

Cognitive processes operate mostly on highly processed sensory data containing ecologically important invariant information, like color.

Conscious perception of the brain activity may be induced externally by signals from senses or by the internal processes, creating persistent distributed integrated activity that can be distinguished from random fluctuations.



Redness, sweetness, itching, pain ... arise due to the physical activations of specific brain areas interpreted by other brain areas.

Brains and computers

Brains: neurodynamics, continuously changing

activation of the brain in space and time.

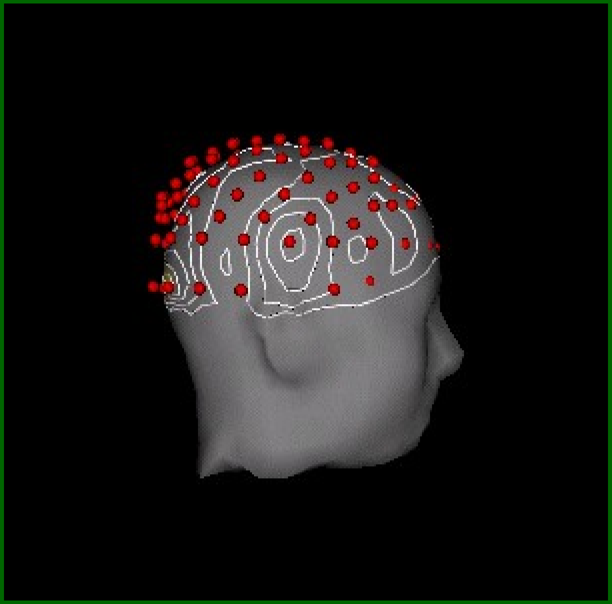
Computer registers: no space, time irrelevant, counting bits in central processors.

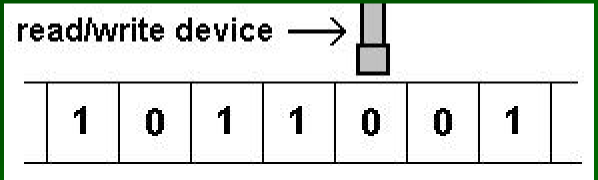
Brain states: distributed neurodynamics, each brain state partially contains in itself many associations, relations, other states.

Mind states are internal interpretations of attractor states.

Computers and robots do not have an equivalent of neurodynamics, nothing similar to attractor states. Analog neurochips may form such dynamics.

W. Duch, J. Minds and Behavior 2005





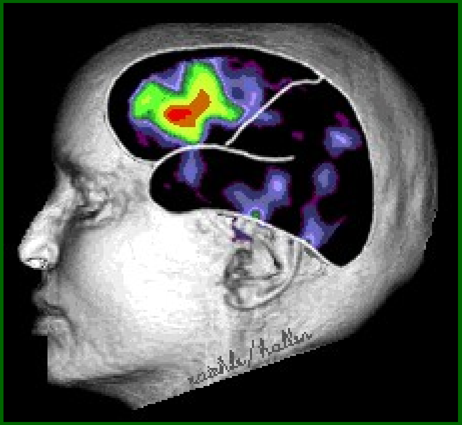
1 0 1 1 0 0 1

Geometric model of mind

Objective ⇔ Subjective.

Brain ⇔ Mind.

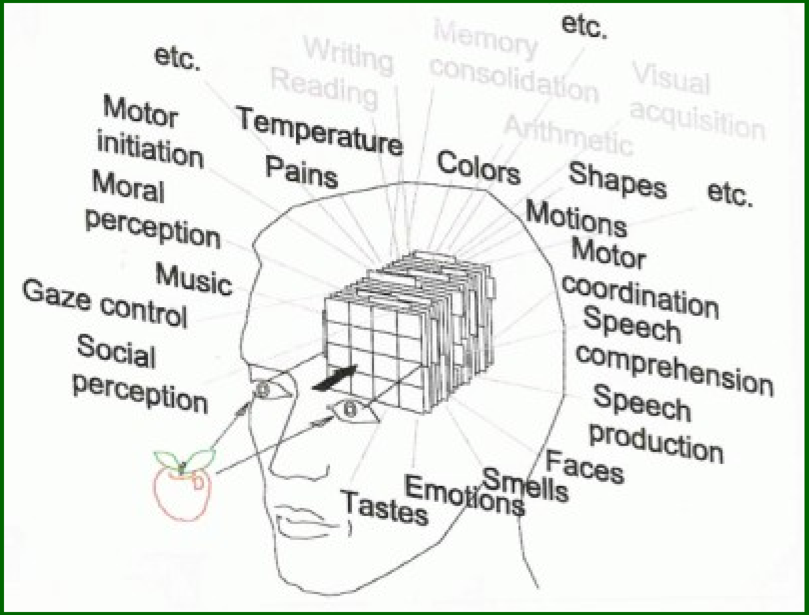
Neurodynamics describes neural activity that can be measured using such neuroimaging techniques as EEG, ERP, MEG, NIRS-OT, PET, fMRI …



How can we describe mental states? Specifying psychological space based on dimensions that represent qualities of experience.

Problem: lack of good phenomenology (E. Schwitzgabel, Perplexities of Consciousness, MIT 2011).

Unusual brains states (drugs, dreams, TMS) induce strange experiences, imagery, hallucinations.

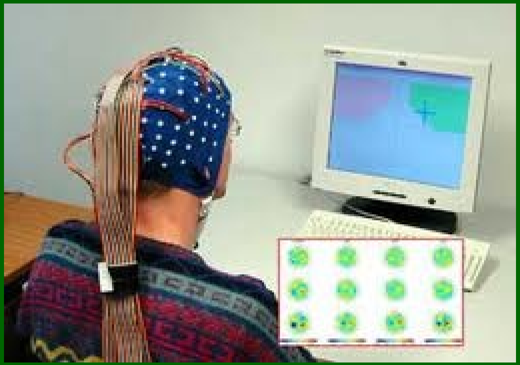


Brain-computer interfaces

Mind reading is an exciting and rapidly developing field.

Brain-computer interfaces (BCI) read and interpret the activity of the brain.

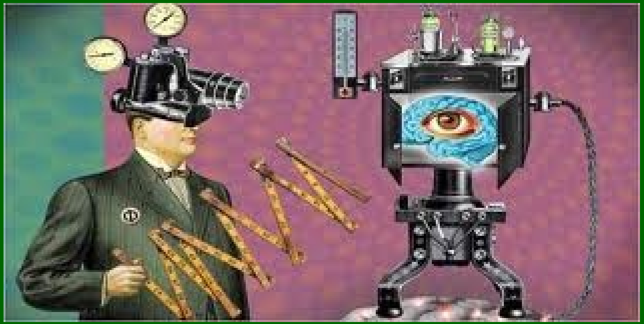
Conscious, intentional activity is detected.

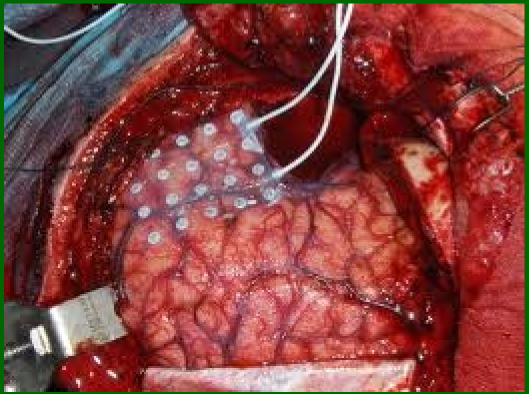


BCI development is motivated by the desire to communicate with people in locked-in or minimal consciousness states (and games -;).

Can we detect signs of consciousness in the same way in artificial brains?

Can we communicate creating resonance states coupling human-robot brains?





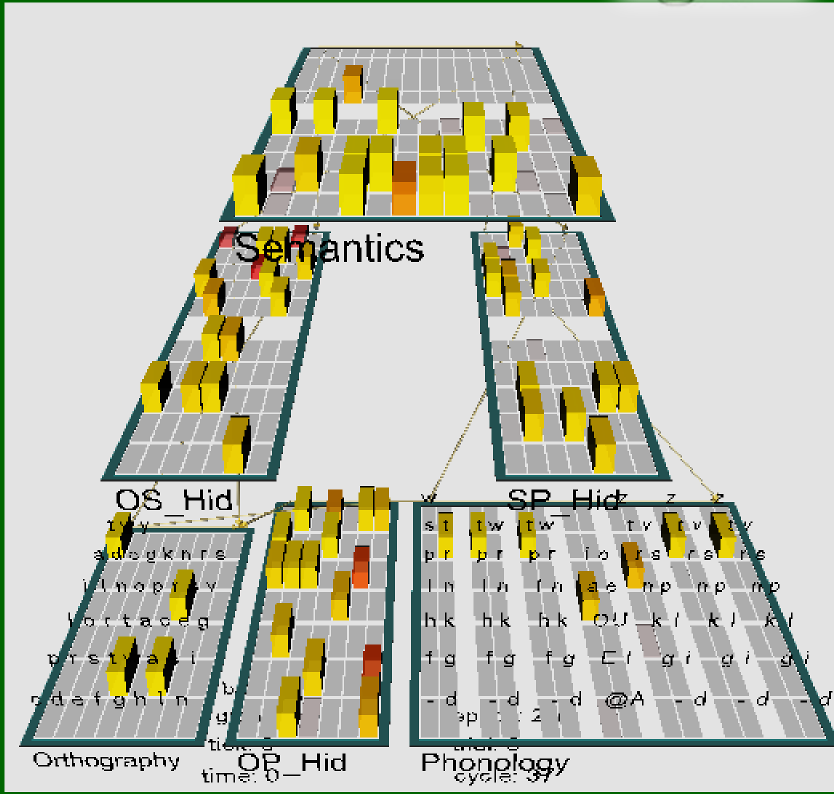
Model of reading & dyslexia

Emergent neural simulator:

B. Aisa, B. Mingus, R. O'Reilly, The emergent neural modeling system. Neural Networks, 2008.

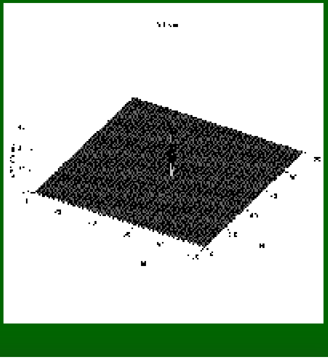
3-layer model of reading:

Recurrent neural network (RNN) with orthography, phonology, and semantic layer = activity of 140 microfeatures that define concepts by distribution of their activations.



Word (written or spoken) presentation => activate semantics, quickly reaching specific configuration of fluctuating active units ⇔ attractor representing concept. Transition to related attractor soon follows. Sequence of such states can be labeled by the activity of phonological or orthographical layers, stream of verbal comments on internal state.

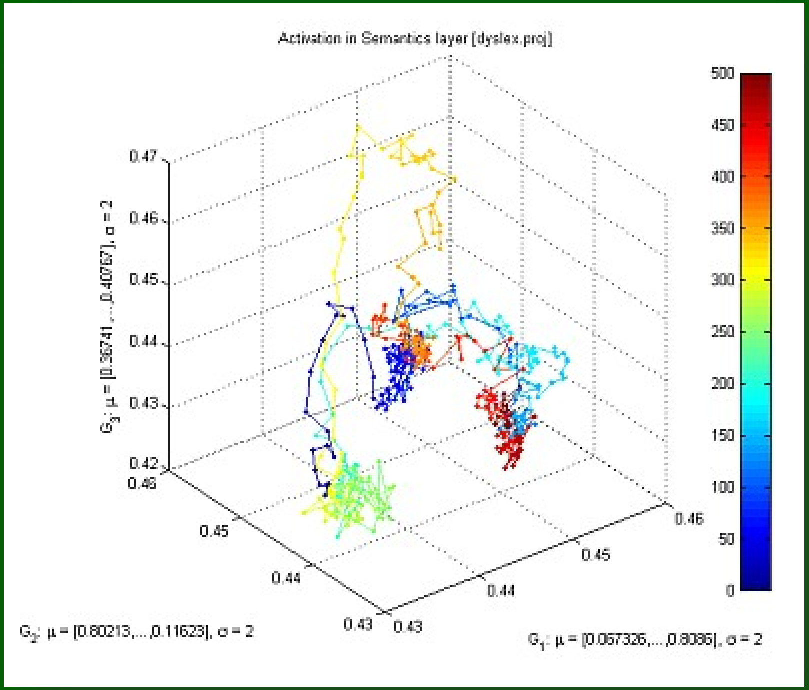
Basins of attractors



Groups of neurons synchronize, become highly active, these activations fluctuate around some specific distributions, inhibiting competing groups of neurons.

Normal case: relatively large, easy associations, fast transitions from one basin of attraction to another, creating “stream of consciousness”.

Brain has about 3 mln minicolumns in the cortex alone, corresponding to units in computational model, so this is a huge space. Here point ⇔ 140D vector.



Basins of attractors = available mental states that can be categorized and identified. They shrink and vanish as neurons desynchronize due to the fatigue; this allows other neurons to synchronize, leading to new mental states (thoughts).

Darwin/Nomad robots

G. Edelman et al, created a series of "noetic brain-based devices" whose behavior is controlled by a simulated nervous system. Principles:

(i) The device engages in a behavioral task.

(ii) The device's behavior is controlled by a simulated nervous system, its design reflects the brain's architecture and dynamics.

(iii) Behavior is modified by a reward or value system that signals the salience of environmental cues to its nervous system.

(iv) The device is situated in the real world.

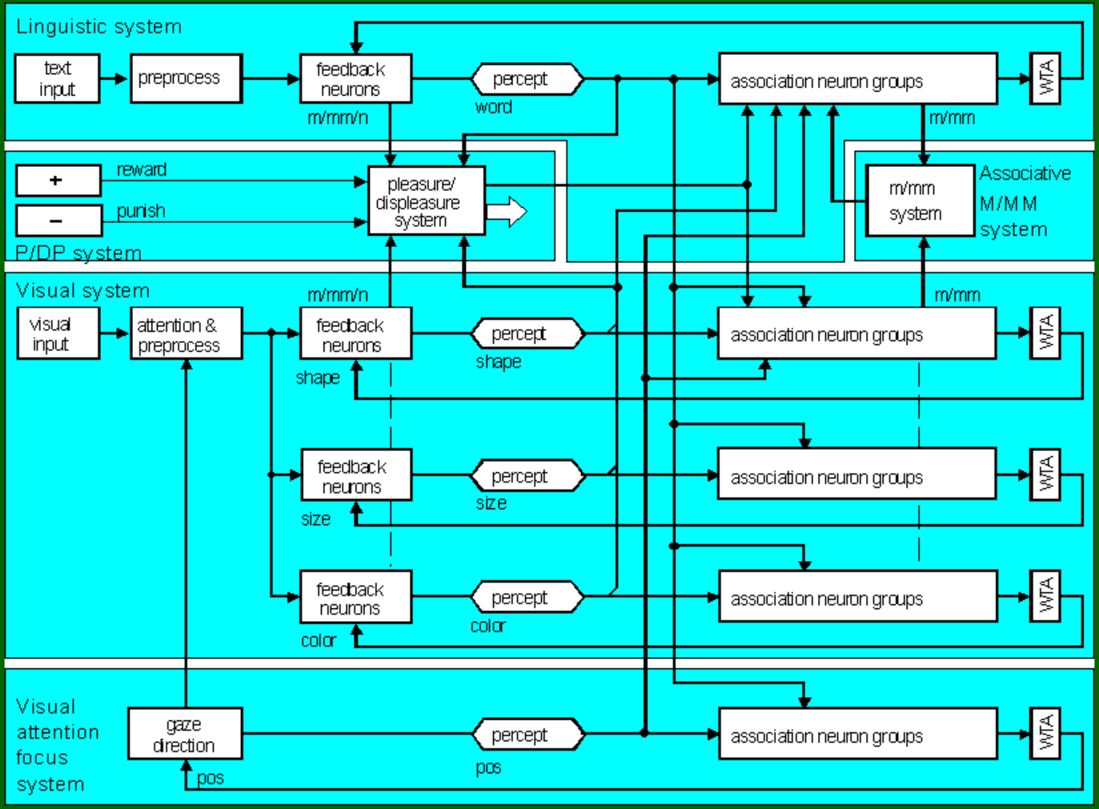
53K mean firing +phase neurons, 1.7 M synapses, 28 brain areas.

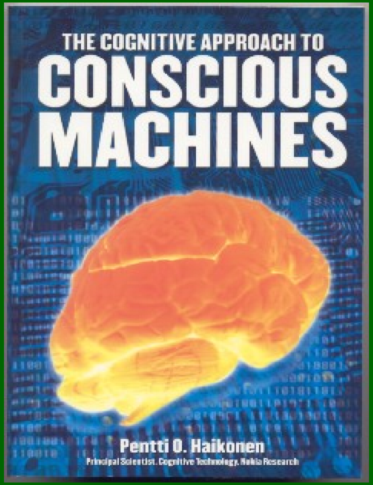
Darwin VII has a mobile base, CCD camera & IR sensor for vision, microphones for hearing, conductivity sensors for taste, and effectors for movement of its base, of its head, and of a gripping manipulator.

All behaviors are emergent and learned, resulting from general principles.

Conscious machines

Many attempts to create brain-inspired cognitive architecture (BICA) are under way. For example, Haikonen has done some simulations based on a rather straightforward design, with neural models feeding the sensory information (with the Winner-Takes-All associative memory) into the associative "working memory" circuits. Such architecture could have interesting neurodynamics.




THE COGNITIVE APPROACH TO
CONSCIOUS MACHINES
Pentti O. Haikonen
Principal Scientist, Cognitive Technology, Nokia Research

Hector, conscious insect

Holk Cruse, Malte Schilling, Mental States as Emergent Properties. From Walking to Consciousness. In T. Metzinger, ed. Open MIND Project 2015.

Hector: insect that walks, plans its path, imagines alternative actions.

A number of higher-level mental states may be attributed to the control system of Hector. “Inner mental states” include intentions, goal-directed behavior guiding robot actions (find food = charging station).

Body properties are coupled with the environment and used in internal model for planning actions (second-order embodiment). Emotions are inherent properties of behavior implemented in the control model reaCog based on recurrent neural networks (RNN). Phenomenal aspect of emotions is understood as an emergent property.

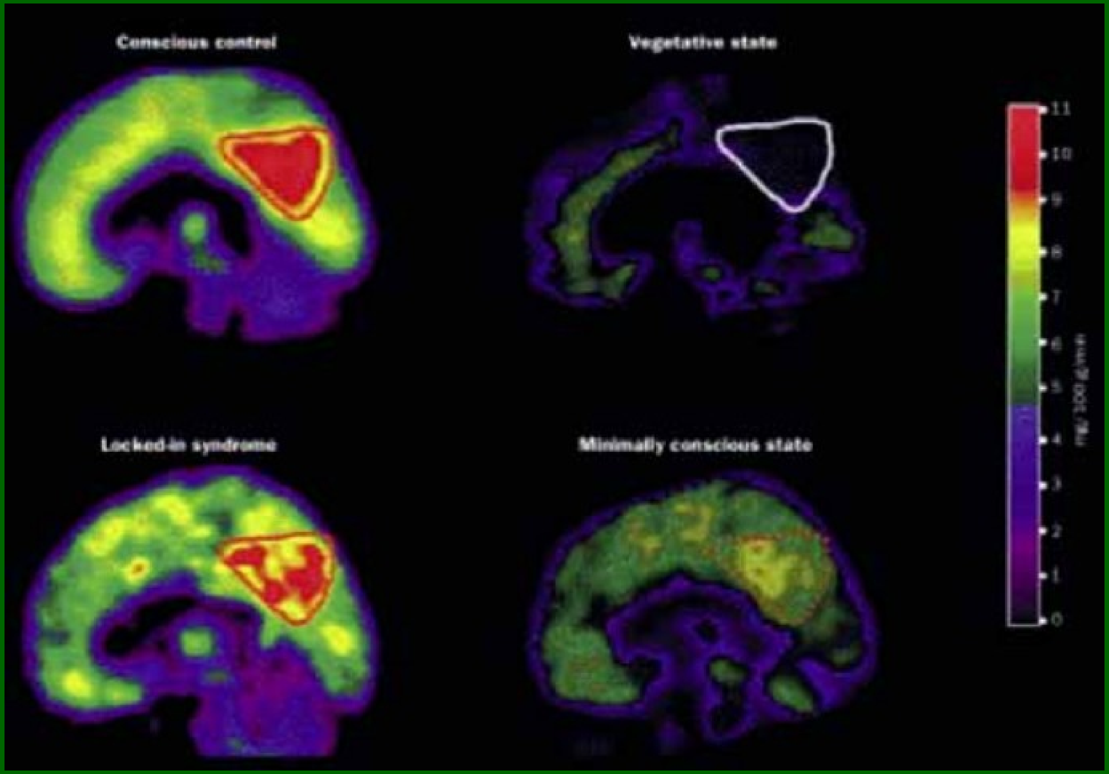
“Depending on its inner mental state, the system may adopt quick, but risky solutions, [… or] take its time to search for a safer solution.” Word units – single word comments.

Measuring consciousness

Neural correlates of consciousness? PET studies show brain activity in normal awake subjects, locked-in subjects, minimal consciousness and vegetative states, and no activity of the dead brain.

Normal consciousness requires distributed integrated brain activity.

Complexity of structure is not sufficient: cerebellum has 80% of all neurons, and no contribution to conscious states.



Laureys S. et al., Lancet Neurology, 2004;3:537-54.

Machine consciousness

Owen Holland (Essex Univ), Tom Troscianko and Ian Gilchrist (Bristol Univ) received 0.5 M£ from the EPSR Council (UK) for a project 'Machine consciousness through internal modeling‘, 2004-2007.

The main focus of interest was strong embodiment in a robot, development of the self-model in Increasingly complex biologically inspired autonomous mobile robots forced to survive in a series of progressively more difficult environments.

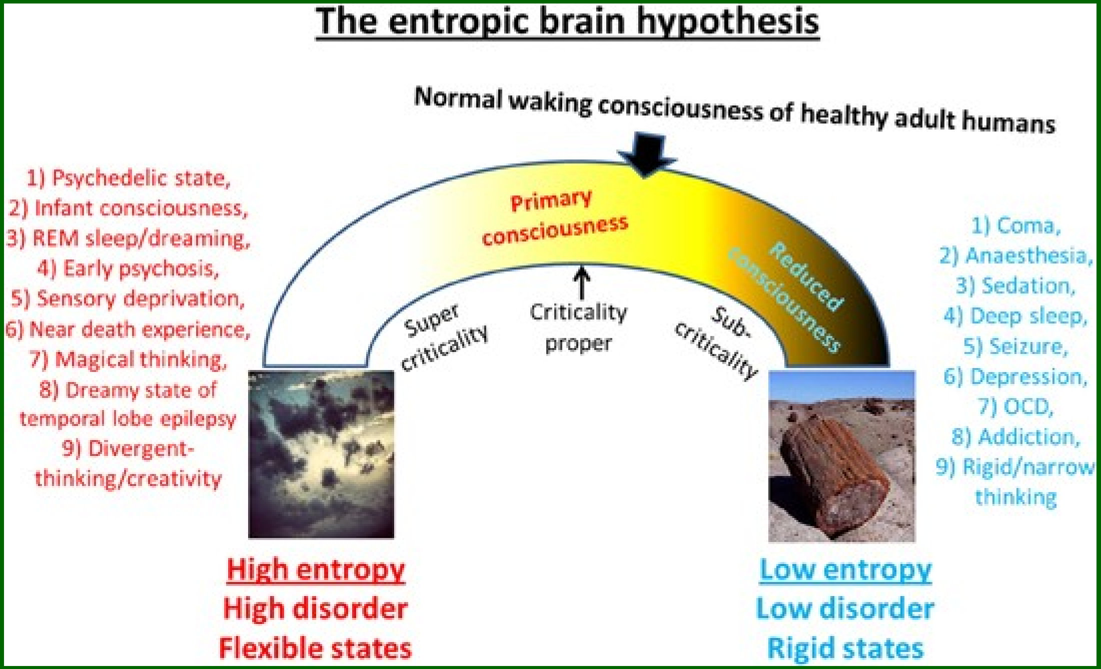
The external and internal behavior of the robots was examined, looking for signs of consciousness.
Still building CRONOS robot.

Measuring consciousness

How to quantitatively measure the level of consciousness in people during anesthesia, epilepsy, coma, disordered states of consciousness, in infants, various animals and machines?

Complexity of neurodynamics: not too chaotic, not too regular.

Several attractor states linking many brain areas, medium entropy.

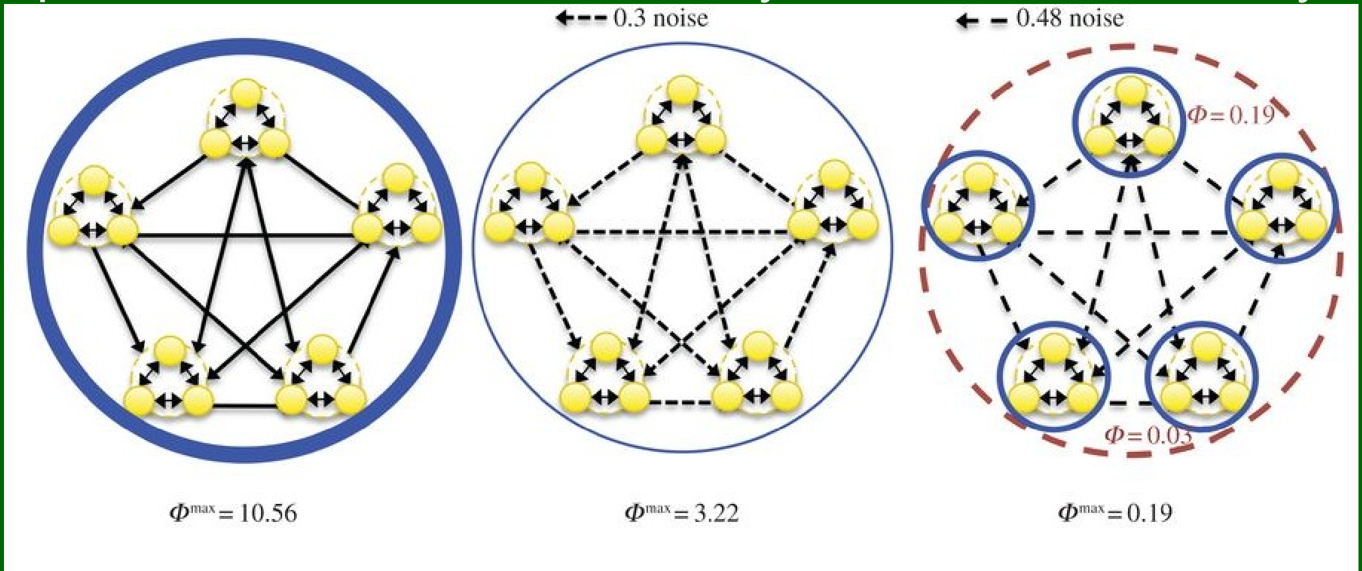


Integrated Information Theory

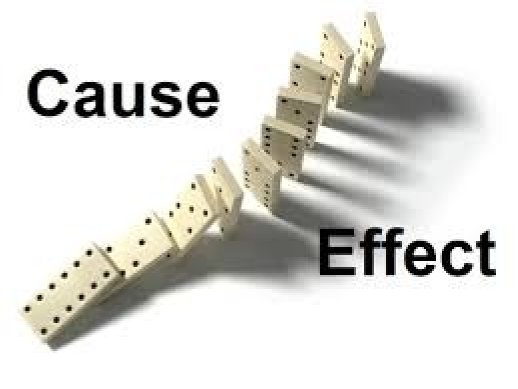
Information integration theory of consciousness (IITC, Tononi, Edelman, Science 1998) defines *integrated information* (F) measure, generated by the neural system, balancing wide integration and information richness.

Seth (2011) proposed causal density, calculated as the fraction of interactions among neural groups that are casually significant.

Tononi, G; Koch, C. (2015). Consciousness: Here, there and everywhere? Phil. Trans. Royal Society London B, 370: 20140167 . Quantity (strength) and quality (shape) of experience is defined by the conceptual structure that is maximally irreducible intrinsically.



IIT postulates



The IIT is based on 5 general postulates, expressed in a rather abstract way below. They may be translated to properties of attractor networks in brain-inspired cognitive architectures.

1. **Intrinsic existence**: must have cause–effect power upon itself.
2. **Structured subsets** of the elementary mechanisms of the system, composed in various combinations, also have cause–effect power.
3. Information is in the cause–effect repertoires is specified by each composition of elements within a system.
4. **The cause–effect structure** specified by the system must be unified: it must be intrinsically irreducible, **a quale**.
5. The cause–effect structure specified by the system must be definite, specified over a single set of elements over which it is maximally irreducible from its intrinsic perspective.

IIT conclusions

Consciousness is a fundamental property of certain physical systems, like brains, having real cause–effect power, specifically the power of shaping the space of possible past and future states in a way that is maximally irreducible intrinsically (Φ measure).

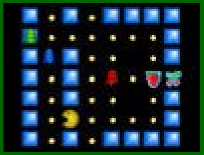
Quantity (strength) and quality (shape) of experience is defined by the conceptual structure that is maximally irreducible intrinsically: quality differs depending on configuration of elements involved.

Feedforward systems cannot be conscious, recurrence is needed.

Computer simulation of the brain are virtual and will not create consciousness. Physical activity of computer elements is not sufficiently integrated in a unified process, breaks down into many mini-complexes of low Φ^{max}.

However, Tononi and Koch do not mention neurocomputers based on massively parallel neurochips (as for ex. in the SYNAPSE project). According to IIT such systems could become conscious and it can be measured.

Final conclusions



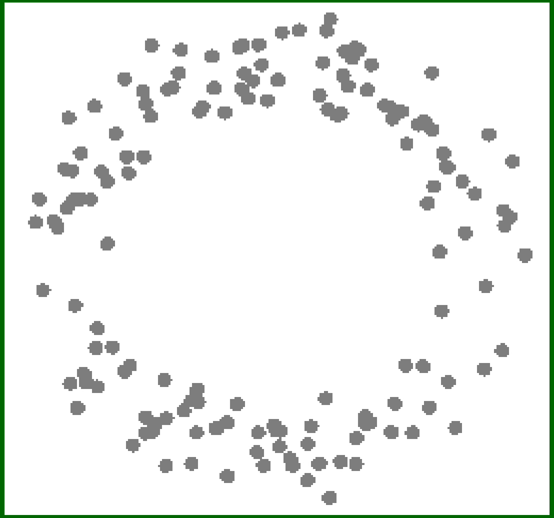
Robots and avatars will slowly converge towards realistic human-like behavior. Will they be conscious? Think about progress in computer graphics.

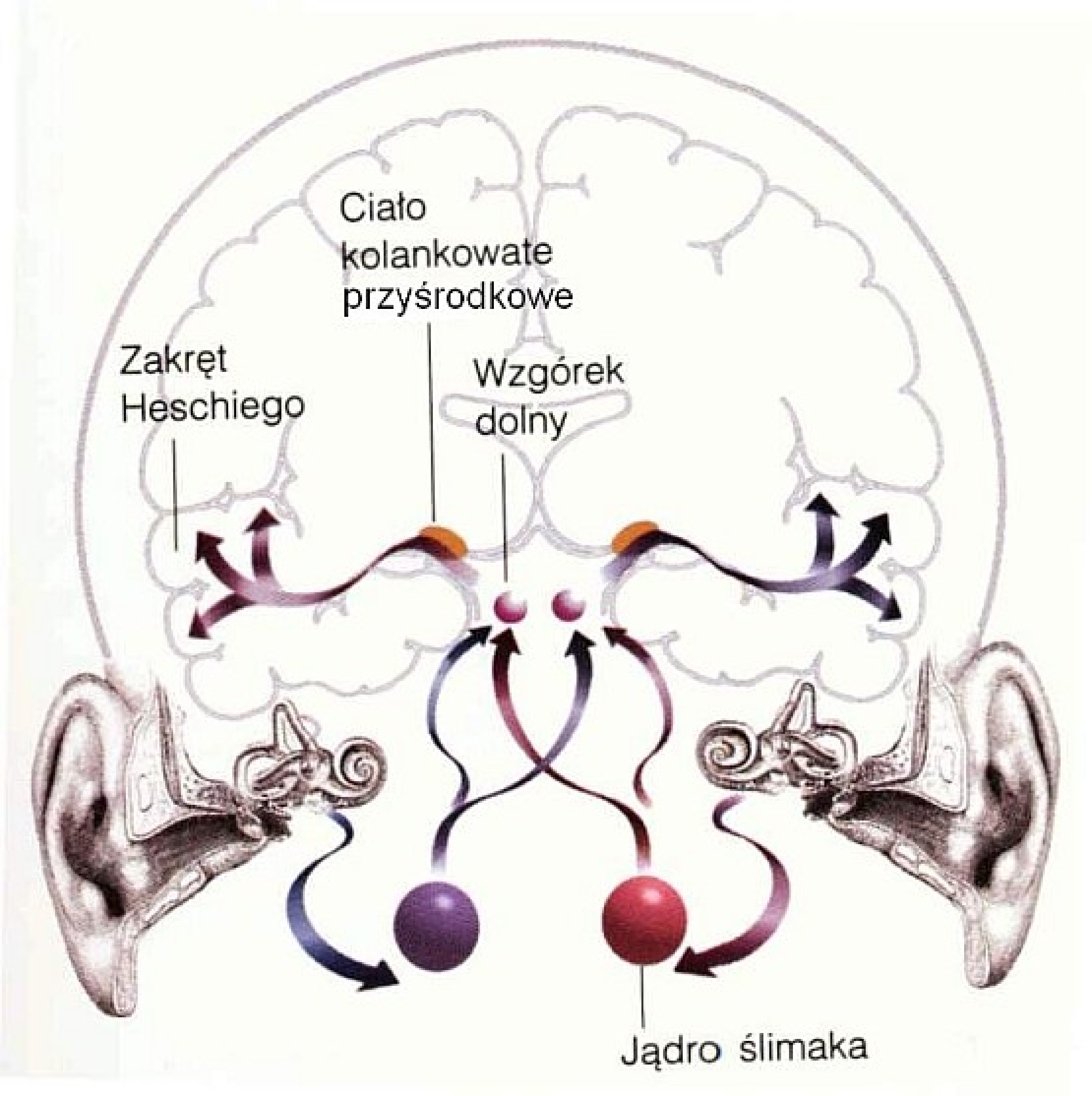
- There are no good arguments against convergence of the neural modeling process in embodied systems and brain-like structure to conscious artifacts.
- Artificial minds of brain-like systems will have to claim qualia; they will be as real in artificial systems as they are in our brains.
- Measures of the level of consciousness based on integrated information theory or its variants will be increasingly useful in medicine and AI.

Creation of conscious artilects will open Pandora's box

What should be their status?
Will it degrade our own dignity?
Is switching off a conscious robot a form of killing?
...
Will they ever turn against us ... or is the governor of California already one of them ?

Thank you for synchronizing your neurons.





Google: W. Duch

Papers, talks, lectures …