Maryland Metacognition Seminar

TOWARD ROBOT CONSCIOUSNESS

ANTONIO CHELLA
University of Palermo
Motivations for MC

- Artefacts like us: consciousness, emotion and affect, experience, imagination, creativity (Robotics)
- Studying natural systems with computer laboratory models (Cognitive Science)
- Proficient machines (Intelligent Control)
When machines will be conscious?

- A conscious machine could require the same complexity of the human brain
- We could be able to build such a machine in 2029
- Human Brain Project (EU Flagship Project)
- Randal Koene *Carboncopies*
- ....
A Brief History

- Nemes: Cybernetic Machines
- Published in Budapest in 1962
- Translated and published in English in 1970
Fig. 137. A preliminary conceptual sketch of the “first person”
1992: Igor Aleksander ICANN Brighton

- Modern view of machine consciousness
- “The hunting season of machine consciousness is open”
“we know of no fundamental law or principle operating in this universe that forbids the existence of subjective feelings in artefacts designed or evolved by humans.”
(C. Koch, concluding comments in final report)
Modern technology and the advancement of artificial intelligence are revolutionizing the way we interact with the world. The concept of machine consciousness is increasingly becoming a reality as we design smarter and more advanced machines. This article explores the implications of machine consciousness, questioning whether these machines will let us live forever or render us obsolete.
EXPERIENCE OF REDNESS
Consciousness and Computational Mind

- The elements of the conscious awareness are caused by information and processes of the computational mind that:
  - are active
  - have other privileged properties

Jackendoff 1987
Computational models for MC

- Consciousness as Information Integration
- Consciousness as Introspection/Monitoring
- Consciousness as Internal Model

http://www.nature.com/ki/journal/v62/n5/fig_tab/4493262f1.html
Information Integration Theory (Tononi)

- Conscious experience is differentiated
  - the potential repertoire of different conscious states is huge
- Conscious experience is integrated
  - every conscious state is experienced as a single entity
- The substrate of conscious experience must be an integrated entity able to differentiate among an enormously big repertoire of different states
Galileo and a photodiode in front of a flashing screen
The same answers!
But... Galileo is able to discriminate among a huge number of states
How much information is generated:
Entropy

\[ H = - \sum p_i \log_2 p_i \]
Integration

- Galileo, a photodiode and a camera in front of a TV screen
- Camera: an immense number of states!
- The camera is a collection of a huge numbers of photodiodes...
- Information not integrated!
Effective Information EI

- S subdivided into two parts A and B
- Perturbation of A: max entropy to A outputs
- EI(A→B): measurements of all the possible responses of B from A
- EI is not symmetric: reverse the procedure for EI(B→A)
The system S can integrate information only if A and B are highly dependent subsets.

- High values of EI(A↔B): strong connections between A and B
- Low values of EI(A↔B): low or no connections between A and B
Φ: Measure for Information Integration

- The bipartition of S for which EI(A↔B) reaches a minimum
- MIP: Minimum information partition
- Φ(S) is the value of EI(A↔B) for MIP
- Φ(S) = EI(MIB A ↔ B)
- Complex: a subset of S with Φ>0 not included in larger subsets with higher Φ
A conscious complex is a complex with high $\Phi(S)$
“Complexes are the subjects of experience, being the locus where information can be integrated”
Consciousness is not an all-or-none but graded by $\Phi(S)$
Complex contributes to conscious experience, the other parts of the systems do not, even if they are connected to it
Experience, e.g., information integration, is a fundamental quantity as mass, charge, energy
Any physical system have subjective experience to the extent that it is capable to integrate information.

It could be possible in principle to build conscious artifact by endowing them with a complex of high \( \Phi(S) \).

A conscious vision machine should be able to differentiate the key features of a scene from the immense range of possible scenes and to integrate them in a detailed description of the scene itself.
Global Workspace Theory (Baars)

- The brain is a collection of unconscious specialized processors
- Consciousness is serial with limited capacity
- Consciousness is associated with a *global workspace* whose contents “broadcast” to many processors
- Contexts shape conscious contents
- Contexts may work together to constraint conscious events
- Motives and emotions are parts of goal contexts
- *Self* is the most general unifying context
Several unconscious processors compete for access to GW in order to recruit more processors.
The winner processor gain GW, i.e., consciousness and broadcast to the other processors
Context may allow for a coalition of processors in order to shape the content of consciousness.
Self is the deepest level of context: the basic intentions and expectations we have towards the world, ourself and each others (Baars)
GWT Implementations

- LIDA (Franklin et al.)
- Shanahan’s Cognitive Architecture
- Dehaene’s Neuronal Workspace Model
- CERA – CRANIUM (Arrabales et al.)
Introspection/monitoring models

- Hierarchy of modules in the computational minds
- Low level modules related with reactive input-outputs
- High level modules related with deliberative planning, reasoning, ...
- Monitor modules
Recursion models

- Recursion of modules in the computational mind
- Level n comprises level (n-1)
- Introspection, self reflection modules
“Mechanized” formal reasoning
Simulation structure: the interpretation model
Association of analogue representation to the symbolic formalism
Exploiting meta-level representations
Reflection about the system itself and its own capabilities
Self-model (McDermott)

- Normal access: information about the world
- Introspective access: information about the robot itself
- Self-model: S=M
- Robot with a model of itself
Layers of reflection (Minsky)

- Multiple layers of critics
- Each layer reflects and critics upon the layers beneath
- Capabilities, limitations and improvements
- EM-ONE (Singh)
Cog-Aff (Sloman)

- Three main levels
- A framework architecture
- Reactive mechanisms
- Deliberative reasoning
- Meta management (reflective processes)
Mental Situation Calculus (McCarthy)

- Introspective reasoning
- Propositions are of mental nature
- A robot may reason about its own mental states
- Situation calculus describes the evolution of robot mental states: knowledge, abilities, intentions, past history, ...
- Introspection as problem solving by considering evolutions of mental states and not just evolution of the external word
Introspective knowledge

- Holds(Know(p), S_i)
- Holds(Know(Not(Know(p))), S_i)
- Holds(Know(Not(Know(Telephone(Mike)))))), S_i)
- The robot may search for Mike’s telephone number in the phone book
Examples of mental actions

- Holds(Knows(p), Result(Learn(p), S_i))
- Occurs(Learn(p), S_i) \rightarrow Holds(F(Know(p), S_i))
- After the learning action occurs, the robot will know p in the future

- Occurs(Forget(p), S_i) \rightarrow Holds(F(Not(Know(p))), S_i)
- Occurs(foo, S_i) \rightarrow Occurs(Forget(p), S_i)
- Forget is a side effect of some event foo
Internal models

- An intelligent agent has an internal model of itself and of the external world
- Capability to simulate the external environment and the body actions
- Generation of expectations
- “Small scale model” of external reality (Craik)
- Popperian creatures (Dennett)
Agent and environment

Agent

World

Holland 2007
Agent with internal model of itself
Consciousness arises by the interaction between the internal model of the agent and the internal model of the environment.
General Framework

Controller \rightarrow Robot \rightarrow Comparator \rightarrow Robot/environment simulator
Implementations

- ECCEROBOT (Holland)
- Starfish robot (Bongard)
- Cicero robot (Chella)
Robust Machines Through Continuous Self-Modeling

Josh Bongard, Victor Zykov, Hod Lipson

Computational Synthesis Laboratory
Sibley School of Mechanical and Aerospace Engineering
Cornell University
The Cicerobot Project

- Cicerobot is a museum guide robot tested in the Archeological Museum of Agrigento.
- Cicerobot allows default and interactive tours
- The user can introduce preferences to plan an ad hoc tour.
- Robot platforms equipped with a pan-tilt stereo head, laser range-finder and sonars.
Cicerobot
Cicerobot II
Robotanic
First order perceptions
(egocentric view)

Second order perceptions
(allocentric view)
ragazzi, museo e Pubblica Amministrazione: un robot per la cultura
ROBOBICO Project

- Joint research with Michele Migliore, Institute of Biophysics, National Research Council, Palermo, Italy.
- Control a humanoid robot by a realistic network of morphologically accurate neurons.
- To learn about the relationships between structure, dynamics, functions and dysfunctions of neuronal circuits.
- To produce experimentally testable predictions facilitating the development of innovative drugs and therapies.
The main goal in this WP is two-fold: to endow robots with more robust methods for coupling sensorial input and action, through realistic neuron models fed with real-world noisy data, and to help the validation of specific hypotheses and predictions under normal and pathological conditions. It will concentrate on the practical interaction between a realistic neuronal network and a commercial anthropomorphic robotic platform. The robot will close the loop between the network activity and the real-world environment, providing the concrete substrate for the experimental testing of both experimental and theoretical findings. To this aim, a low-latency interface between the NEURON simulator running on a cluster and the robot will be devised. Features extracted from the continuous stream of pre-processed perceptual data will be classified into a set of target classes for each perceptual channel (i.e. color, shape, sound, approaching, avoiding) and will be connected to the CA3 inputs of the network. The activity of each neuron will contribute to determine the output pattern of the network, which will drive the behavior of the robot by modulating the activation of predefined parametric motor functions. It should be noted that the accurate (biophysical) modeling of physiological properties of all brain regions subserving the activities needed for the correct operation of the network is well beyond the scope and possibilities of the present proposal. For this reason, some functionalities – in particular those related to the classification of perceptual data and behavior selection – will be addressed by more conventional or state-of-the-art machine learning techniques. Figure 8 schematically depicts the overall idea to be exploited in tight interaction with the modeling and theoretical tasks in the other WPs.
Role of Embodied Interaction for MC

- Collaboration with Hiroshi Ishiguro (Osaka University)
- Essential embodiment
- Motions of head, lips, arms
- Sense of presence
Basic Embodiment for Interactions

Imitation learning

← Prosody →

← Gestures →

User
A Slogan for RoboticsLab

BUILDING ARTIFACTS ABLE TO IMPROVE OUR INNER LIFE
Thank you for your attention!