Self-Awareness:
Emerging views and potential impacts

James L. Crowley
Professor, Grenoble INP
Laboratoire Informatique de Grenoble
INRIA Grenoble Rhône-Alpes Centre de Recherche

Self-Awareness:
Emerging views and potential impacts

Outline

- Awareness
  - Models from Cognitive Psychology
  - Applications for Informatics
- Self-Awareness
  - Models from NeuroBiology
  - Applications for Informatics
- Why study machine self-awareness?
- Potential impacts on technology and society

But first a warning….

We are all incompetent….

The scientific study of the technology for building Autonomous Self-Aware systems is in pre-paradigm.

We do not have:

- A unified scientific community that share problems and problem solutions
- a working theory for how to build such systems.
- Consensus on the fundamental concepts

We do have

- Potential contributions from many established disciplines
- Communities of interested researchers.
- An historic grand challenge that has intrigued humanity for at least 3000 years.

Introduction

We are all incompetent….

The scientific study of the technology for building Autonomous Self-Aware systems is in pre-paradigm.

We do not have:

- A unified scientific community that share problems and problem solutions
- a working theory for how to build such systems.
- Consensus on the fundamental concepts

We do have

- Potential contributions from many established disciplines
- Communities of interested researchers.
- An historic grand challenge that has intrigued humanity for at least 3000 years.

Introduction

Method:

- Explore (a few) concepts and theories from the human sciences: Human factors, cognitive science and NeuroBiology
- Examine how to use these as paradigms for designing artificial systems.

Concepts to examine:

Awareness, Autonomy, Self-Awareness.

Awareness

- Vigilance against danger or difficulty.
- Having knowledge of something.
- The ability to perceive, to feel, or to be conscious of events, objects or sensory patterns.
- Conscious of stimulation, arising from within or from outside the person

Models of awareness have been studied and applied for human factors in aviation since at least 1914.

Human Factors: Mica R. Endsley

Mica Endsley, Ph.D., P.E.

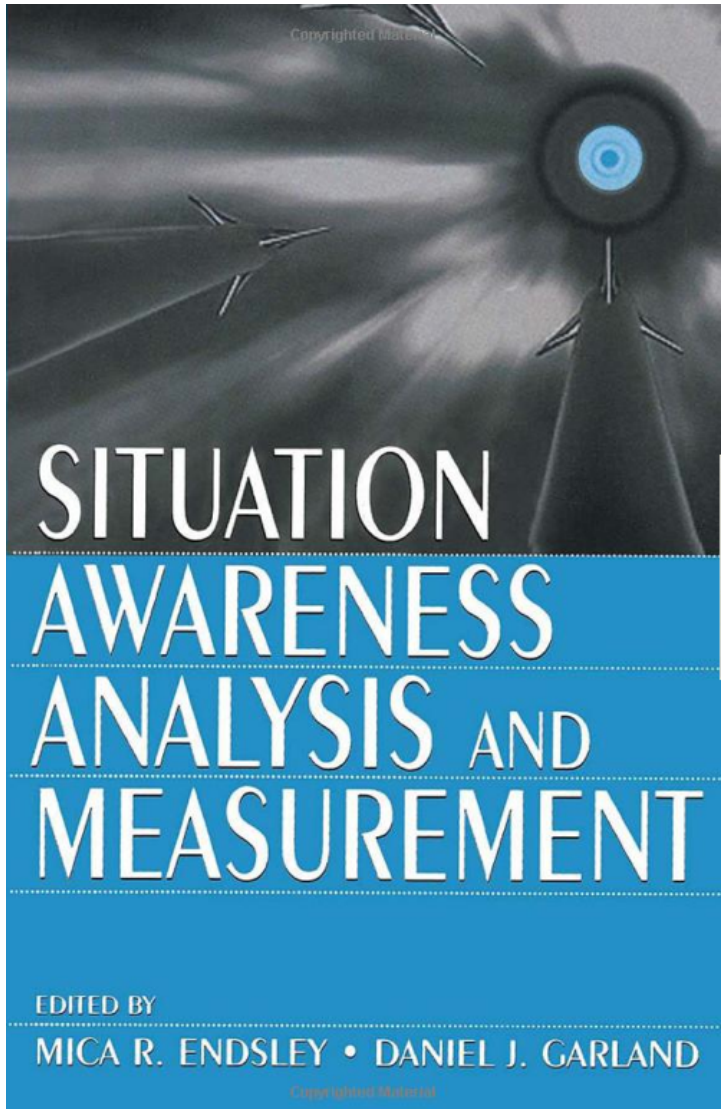
PhD USC 1990

editor-in-chief of the Journal of Cognitive Engineering and Decision Making

President: SA Technologies

Specialty: Cognitive Engineering

Application Domain: Aviation and critical systems.



Human Factors: Mica R. Endsley

Method: Task analysis using Situation Awareness Model

Key Publications:

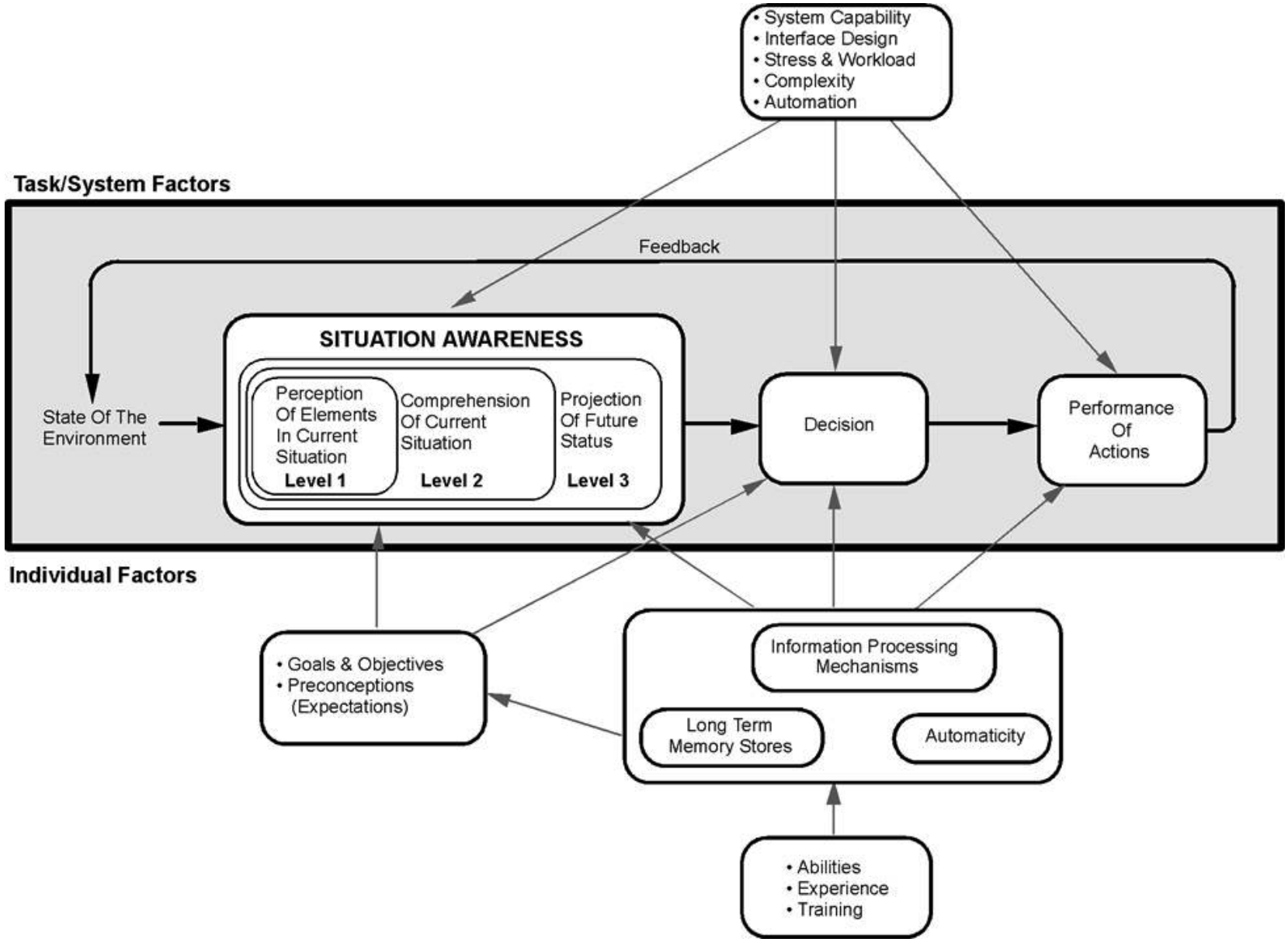
MR Endsley, DJ Garland, (2000), Situation awareness: analysis and measurement, Lawrence Erlaum Associates.

Endsley, M.R., (2000), Theoretical underpinnings of situation awareness: A critical review, Ch1, Situation awareness analysis and measurement, pp3-32.

M.R. Endsley, B. Bolte, D.G. Jones (2003), "Designing for situation awareness: An approach to user-centered design", Taylor & Francis, New York,

Endsley, M. R. (1995b). Toward a theory of situation awareness in dynamic systems. Human Factors 37(1), pp 32-64.

Endsley, M.R., (1995), Measurement of situation awareness in dynamic systems, Human Factors: The Journal of the Human Factors and Ergonomics Society, vol 37, no 1, pp 65-84, 1995.



Endsley's Model of the Dynamic Decision Making Process.
Figure 2 from (Endsley 2000)

Levels in Situation Awareness

Situation Awareness (Endsley): The Perception of [relevant] elements of the environment in a volume of space and time, the comprehension o their meaning and the projection of their status in the near future.

Levels in Situation Awareness

1: Sensing: Sensing of entities relevant to task

2: Assimilation: association of percepts with models that predict and explain.

3: Projection: Forecast events and dynamics of entities

To use Endsley's model for artificial systems we need a programmable theory for perception, assimilation and projection.

Cognitive Science: Philip N. Johnson-Laird

Philip N. Johnson-Laird

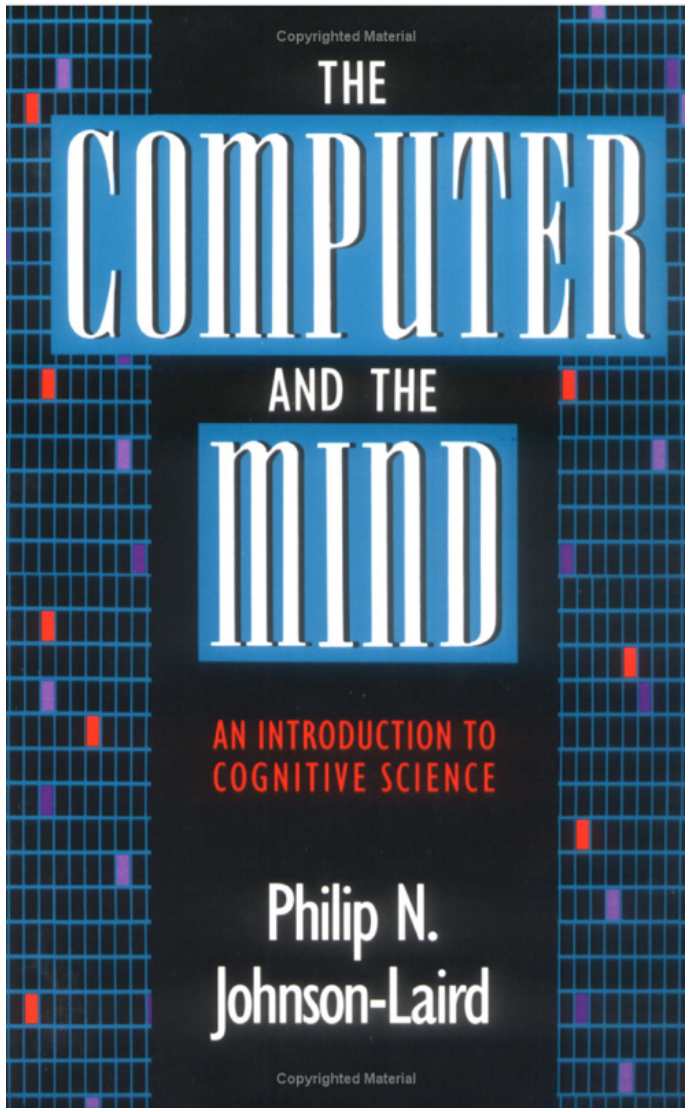
PhD Psychology, 1967, University College London

Stuart Professor of Psychology at Princeton Univ.

1971-1973: Inst. of Advanced Study, Princeton U.

1973-1989: Laboratory of Exp. Psychology, Univ of Sussex

1989- Applied Psychology Unit, Princeton Univ.



Cognitive Science: Philip N. Johnson-Laird

Research Method:

Situation Models as theory of mental models for natural language and inference.

Publications:

Johnson-Laird, Philip N . How We Reason. Oxford University Press. 2006

Johnson-Laird, Philip N . Computer and the Mind: An Introduction to Cognitive Science. Harvard University Press. 1998

Johnson-Laird, Philip N with Ruth M. J. Byrne. Deduction. Psychology Press, 1991

Johnson-Laird, Philip N. Mental Models: Toward a Cognitive Science of Language, Inference and Consciousness. Harvard University Press, 1983

Situation Models:
A model for human cognitive abilities

P. Johnson-Laird 1983 – Mental Models.

A model to describe human cognitive ability to

1) Provide context for story understanding
2) Interpret ambiguous or misleading perceptions.
3) Reason with default information
4) Focus attention in problem solving

Situation Models:
A model for human cognitive abilities

Application Domains in Cognitive Psychology

- Understand spoken narration
- Understanding text and literature
- Spatial reasoning
- Social interaction
- Game playing strategies
- Controlling behavior

Situation Models: Concepts

P. Johnson-Laird 1983 – Mental Models

Situation: Relations between entities

Entities: People and things. Correlated observations.

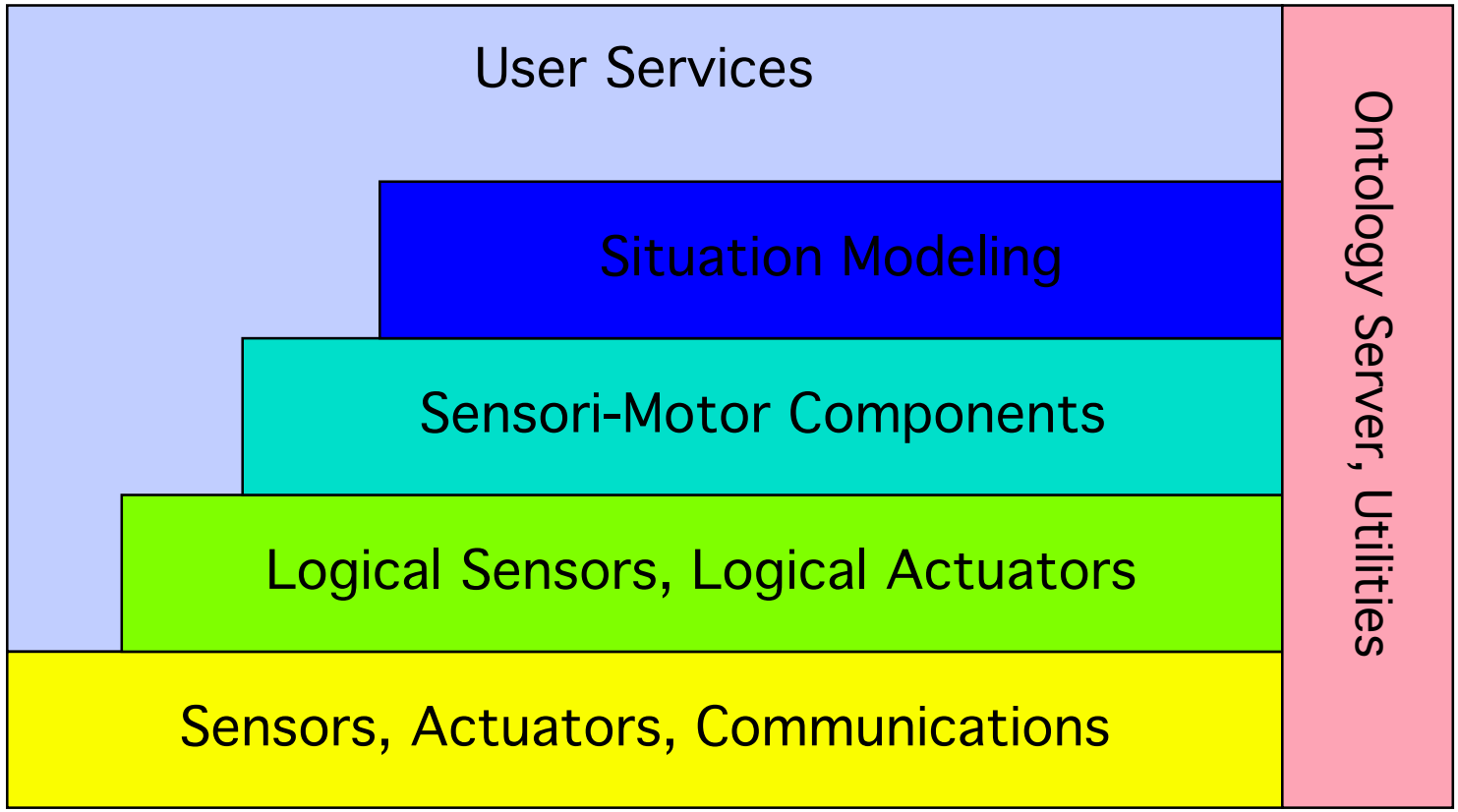
Relations: An N-ary predicate (N=0,1,2…)

Examples: John points-at board, John looks-at audience

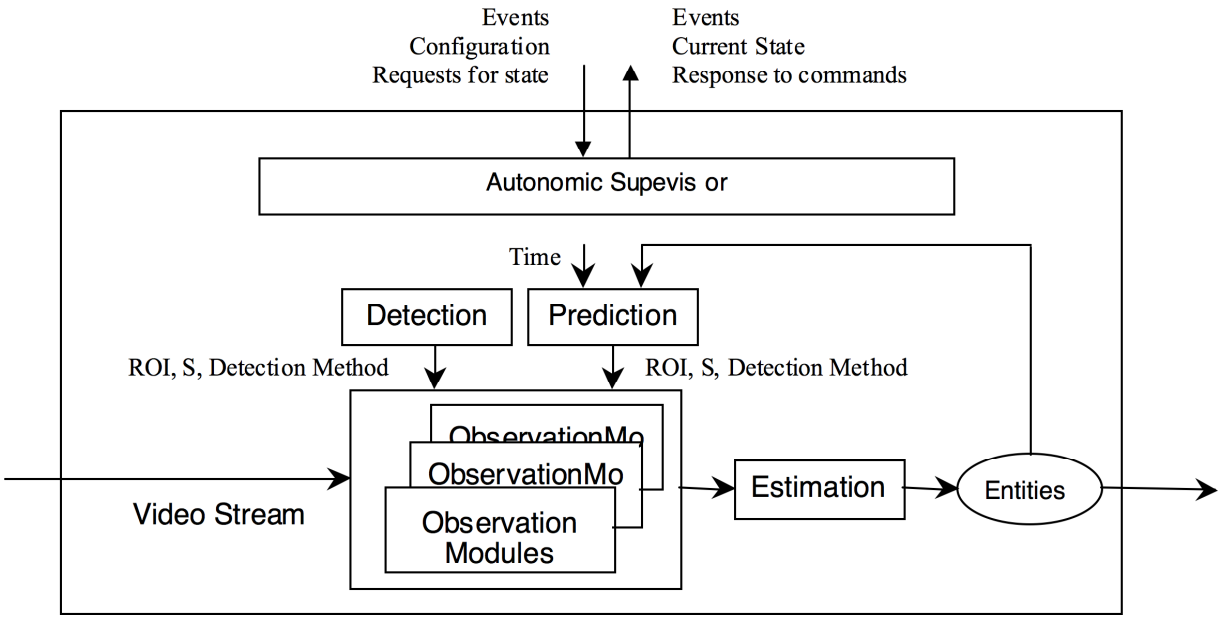
Situation Models can be used as a programming theory.

Software Architectural Reference Model

(ICT CHIL - Computers in the Human Interaction Loop)



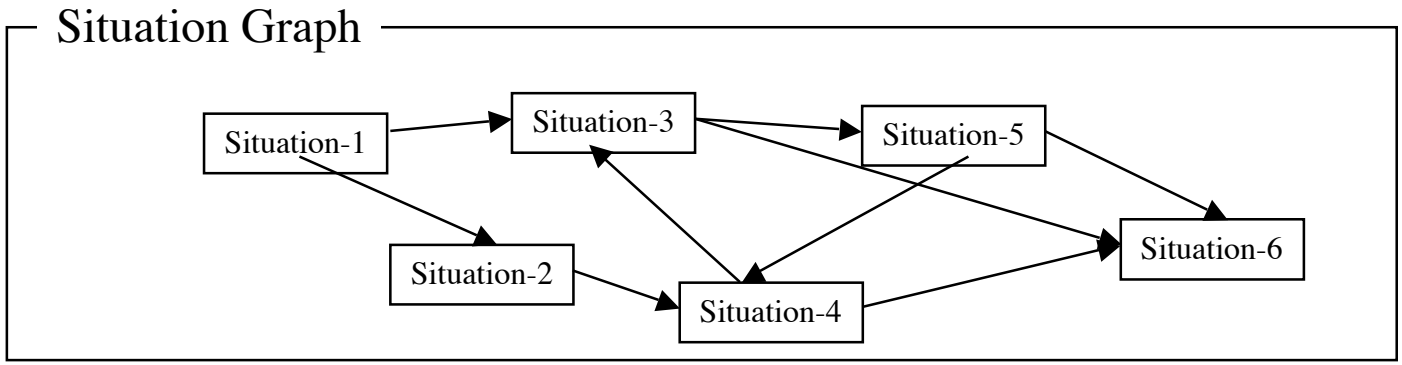
Assimilation: Perceptual Component



Autonomic Supervisor Provides:

- Module execution
- Parameter Regulation
- Interpretion for external communication
- Description of State and Capabilities

Projection: Situation Graph



A situation graph describes a state space of situations

Each Situation provides information for:

- Focus of Attention: entities and relations for the system to observe
- Default information (Context)
- Predictions about possible next situation
- System actions (prescribed, allowed or forbidden)

Situation Models:
A Tool for Observing and Understanding Activity

Examples of applications of situation models

- Lecture Recording system
- Meeting event recording system
- Automatic Privacy filter (for mediaspace)
- Monitoring activity for assisted living
- Teaching polite interaction to robots
- Automated Video surveillance
- Customer observation systems.

An example:
Observing a card Game

Actors:

- Jerome, Sonia and Stan

Objects:

- Card table, cards

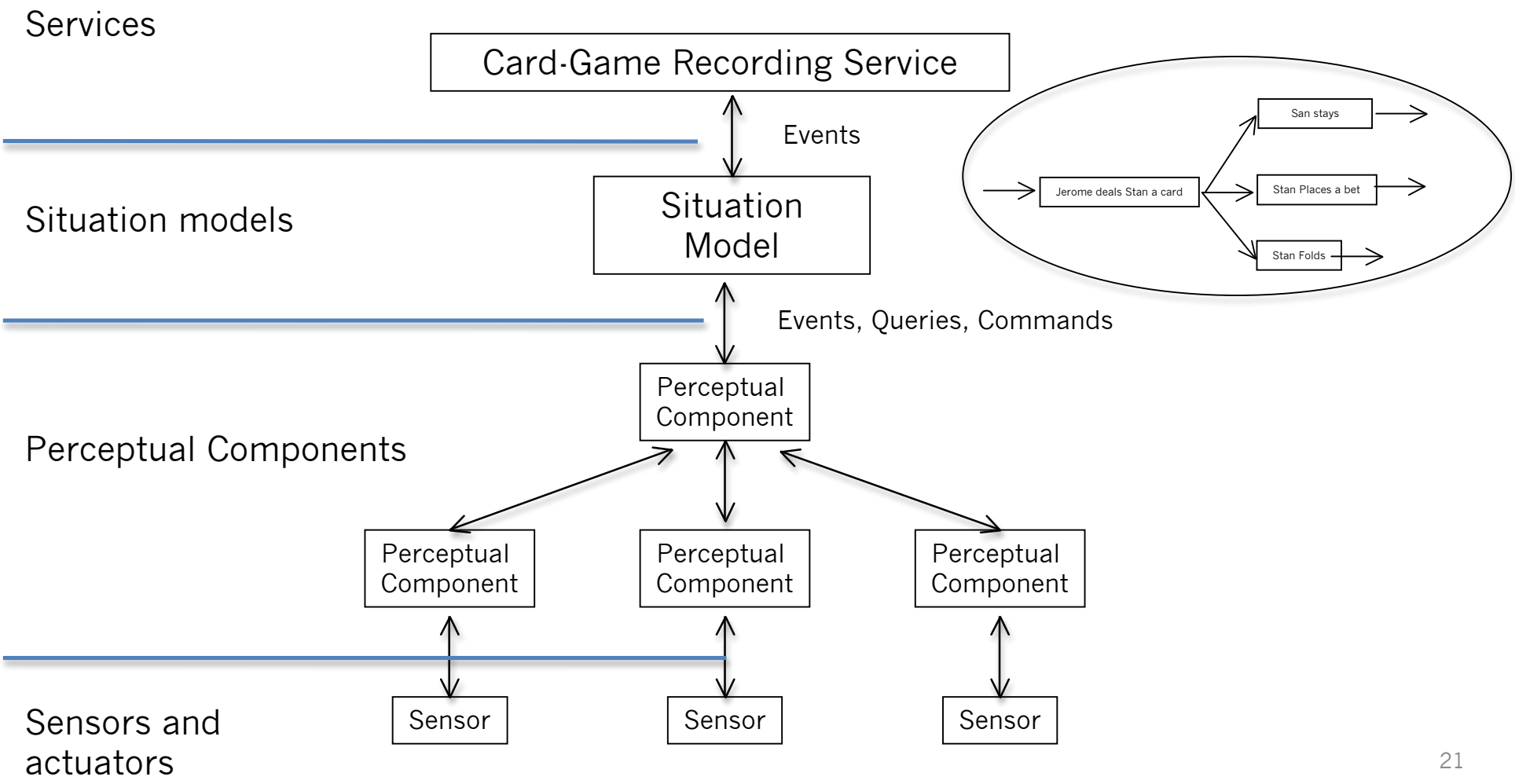
Roles:

- Dealer, player

Relations:

- <Dealer> deals to <player>
- <player> talks to <player>
- <player> Folds
- <player> makes a bet
- <player> talks to <x>

An example:
Recording events in a card Game



Awareness

Machine awareness (in the sense of Endsley or Johnson-Laird) is within the current state of the art in machine perception

Awareness of self requires more: A sense of self.

Awareness of Self

Self-awareness: Aware of oneself as an individual.

Self: Consciousness of one's own identity

Consciousness:

- an alert cognitive state in which a subject is aware of himself and his situation
- knowing and perceiving; having awareness of surroundings and sensations and thoughts;

In medicine, consciousness is assessed by observing a patient's alertness and responsiveness to stimuli

Autonomy and Self

Assertion: a sense of self requires autonomy

Autonomy:

- From auto "self" + nomos, "law": one who gives oneself his/her own law
- Existing as an independent entity
- Self-governing, Self-protecting.
- Able to self-maintain functional integrity.

Autonomous System: An independent system that controls itself and maintains its own integrity

Autonomic Computing provides an enabling technology for Autonomy.

Autonomic Computing

Autonomic computing : a metaphor inspired by natural self-governing systems, and the autonomic nervous system found in mammals.

March 2001 Keynote address to the National Academy of Engineers by Paul Horn (IBM vice president)

Autonomic computing systems are systems that manage themselves given high-level objectives from administrators.

A scientific community for autonomic computing for comptuer operating systems has emerged.

Concepts from autonomic computing are also having impact on Computer Vision, Robotics, and Computer Networks.

Autonomic Nervous System (ANS)

The ANS regulates the homeostasis of physiological functions

The ANS is not consciously controlled.

Commonly divided into three subsystems:

- Sympathetic nervous systems (SNS)) (fight or flight)
- Parasympathetic nervous system (PNS) (rest and digest)
- Enteric nervous systems (ENS) (the second brain)

The ANS maintains internal homeostasis.

Homeostasis

Homeostasis :

- the ability of a system to regulate its internal environment and to maintain a stable, constant condition.Typically used to refer to a living organism.

Defined by Walter Bradford Cannon (1932) from ancient Greek (homos, Similar), (histēmi, Standing still).

Inspired by the milieu interieur desribed by Claude Bernard in 1865.

- Homeostasis is fundamental to life.
- Humans extend homeostasis to their external environment.
- Awareness (Vigilance against danger or difficulty) is part of the extension of homeostasis.

Autonomic Properties for Software Systems

The essence of Autonomic Computing is Self Management.
(Klephart and Chess, 2002).

Self-Configuration:	Automatic configuration of components;
Self-Healing:	Automatic discovery, and correction of faults;
Self-Optimization:	Automatic monitoring and control of resources to ensure the optimal functioning with respect to the defined requirements;
Self-Protection:	Proactive identification and protection from arbitrary attacks.

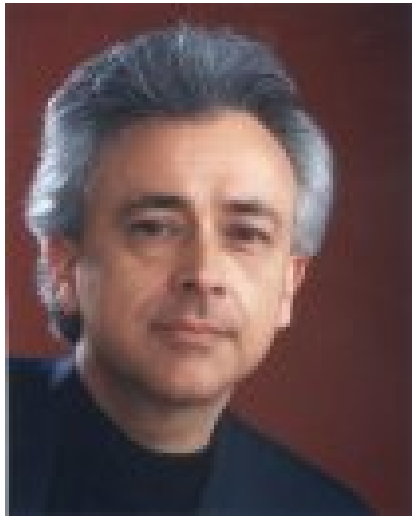
Autonomic Properties for Computer Vision

As used of Perceptual Components in the ICT CHIL project.
(JL Crowley, Autonomic Computer Vision, Int. Conf on Vision Systems, 2007.)

Auto-configuration:	The component can configure its own operating parameters
Auto-regulation:	The component adapts parameters to maintain a desired process state.
Self-description:	The component provides descriptions of the capabilities (to component registry) and the current state of the process (on request).
Self-monitoring:	The component supervisor estimates state and quality of service for each processing cycle.
Self-repair:	The component can detect and correct conditions by reconfiguring modules.

The use of Self in this context is abusive! (pet-peeve).
Systems and services do not (currently) have a sense of self.

Neuroscience: Antonio R. Damasio

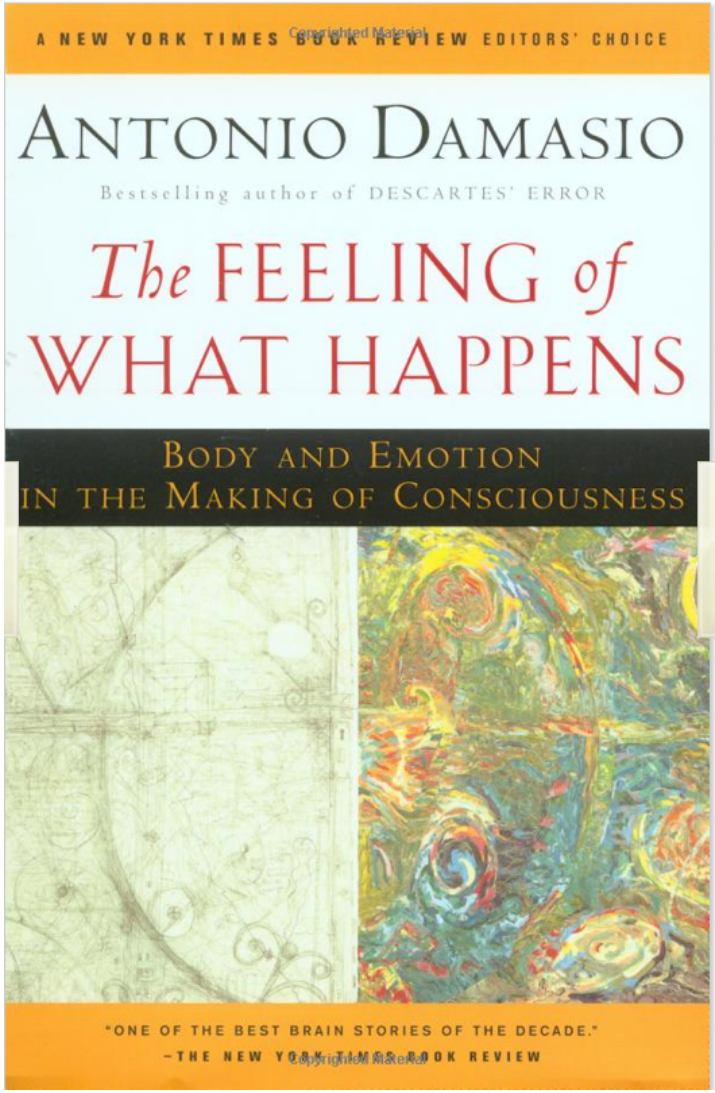


Dr. Antonio Damasio

Professor of Neuroscience USC

USC Brain and Creativity Institute Studies

Research: neurobiology of the mind, especially neural systems for memory, language, emotion, and decision-making.



Neuroscience: Antonio R. Damasio

Research Method:

Correlates loss of cognitive abilities with damage to neural structures

Books :

- Descartes' Error: Emotion, Reason, and the Human Brain 1994
- The Somatic marker hypothesis and the possible functions of the prefrontal cortex, 1996
- **The Feeling of What Happens: Body and Emotion in the Making of Consciousness, 1999**
- Looking for Spinoza: Joy, Sorrow, and the Feeling Brain, 2003

Emotions and Appetites

Somatic marker mechanism (Damasio): Theory of how perceptions interact with emotions.

Emotions regulate homeostasis in reaction to stimuli from the external world.

Homeostatic emotions are feelings evoked by internal body states, which modulate behavior.

Examples: Thirst, hunger, feeling hot or cold (core temperature), feeling sleep deprived, salt hunger and air.

Appetite: An instinctive physical desire, especially one for food or drink. Appetites motivate much of human behaviours

A. Damasio : Some vocabulary

Images: Neural patterns with structure representing sensory stimuli: Visual, auditory, olfactory, gustatory, somatosensory (body).

Images Depict physical properties of entities

May be driven by senses or by internal stimuli (dispositions)

Are dynamic (spatio-temporal).

Are not always conscious. (example: Blind-sense).

Sensori-motor maps: express interaction between an organism and an entity.

Second order maps: represent changes in Sensori-motor maps.

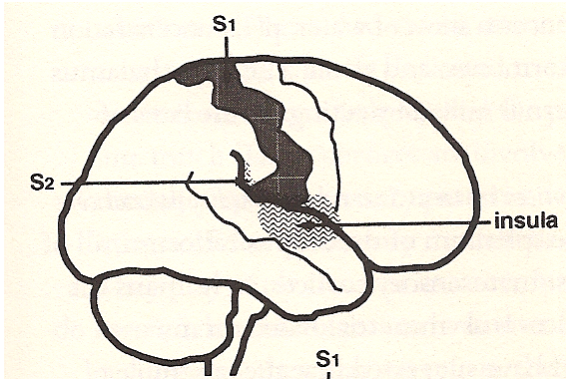
Categories of Self (A. Damasio)

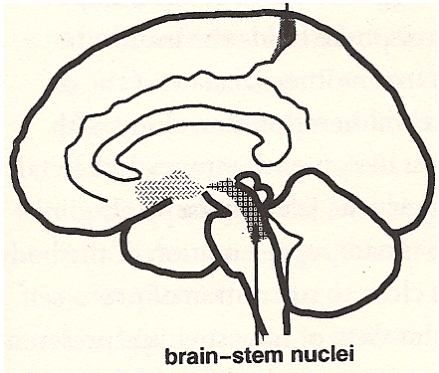
Core-Self: A transient entity; recreated for each object with which the brain interacts

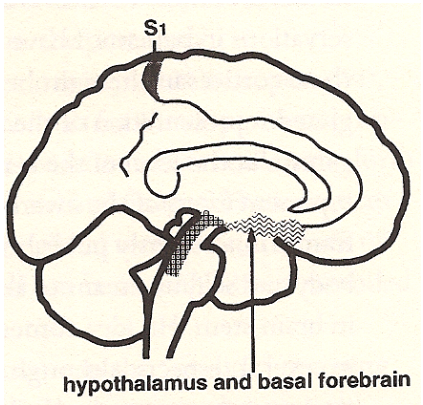
Proto-Self: A representation of core self in a second order neural map (image).

Autobiographic Self: Permanent dispositional records of core-self experiences. Systemized memories of situations in which core consciousness was involved with invariant characteristics of life.

Location of some proto-self structures

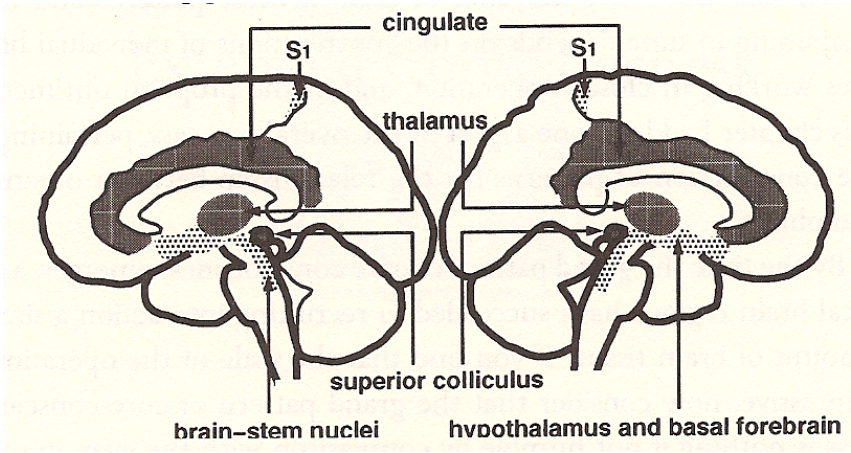






From Damasio, The Feeling of What Happens: Body and Emotion in the Making of Consciousness, 1999, figure 5.1 pp 155

Proto-Self and Second Order Maps



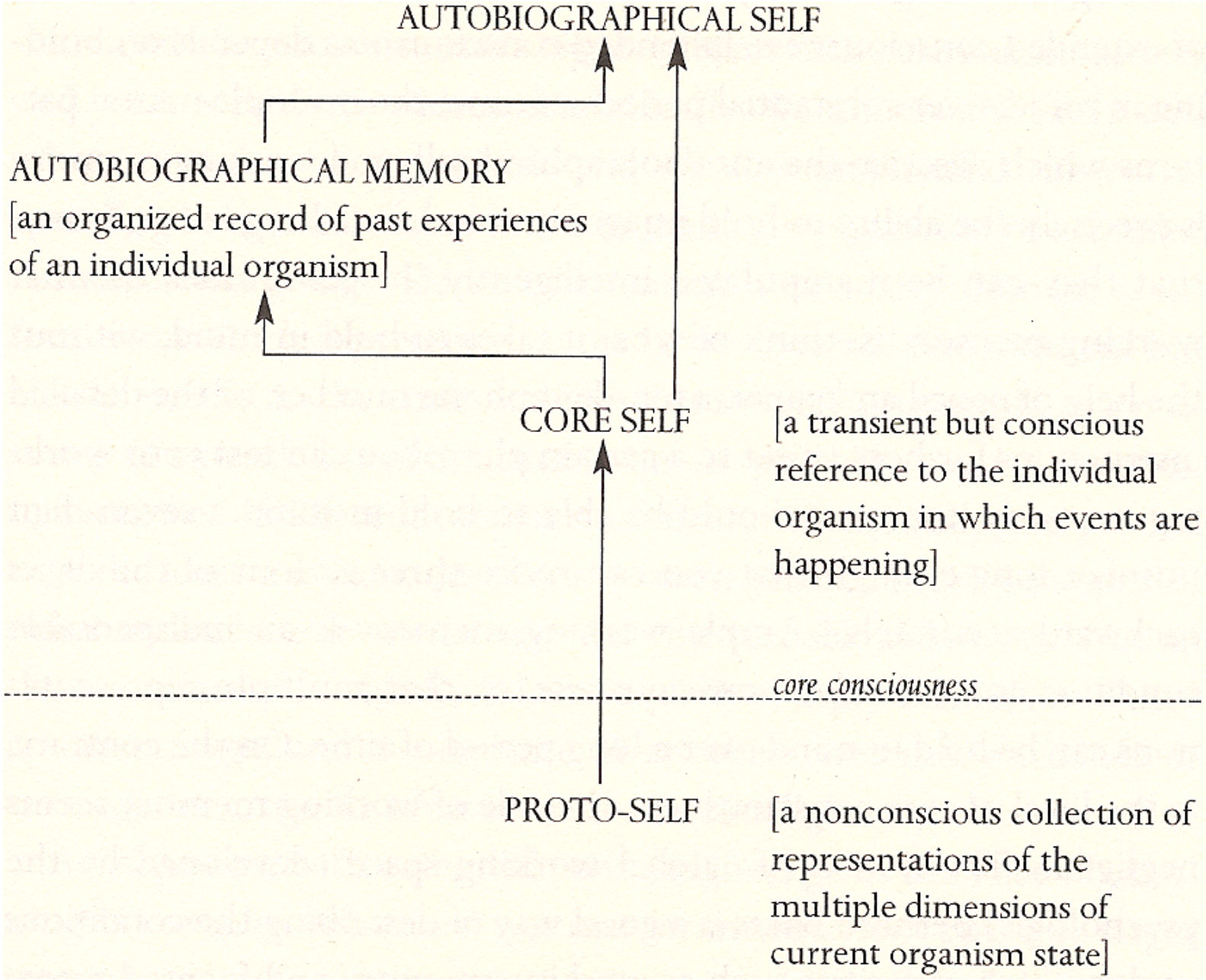
From Damasio, The Feeling of What Happens: Body and Emotion in the Making of Consciousness, 1999, figure 6.3 pp 193

Self Awareness (A. Damasio)

Consciousness

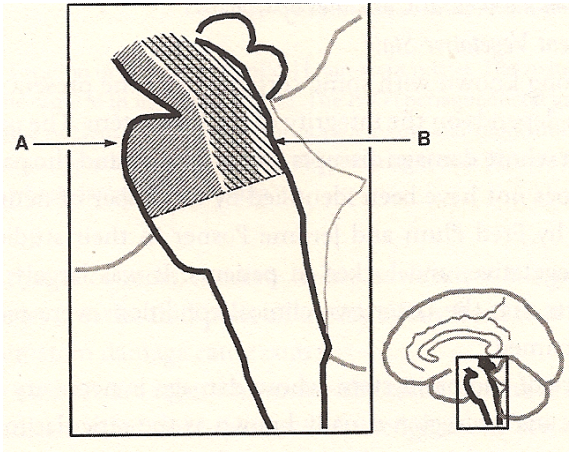
Core Consciousness: an imaged, non-verbal account of the organisms own state.

Extended Consciousness: Awareness of entities and events in relation to autobiographical self.



From A. Damasio (1999),The Feeling of What Happens: Body and Emotion in the Making of Consciousness p199

Neurobiological locus of self-awareness



Damage to structure front brain stem(A) causes locked-in-syndrom: Consciousness with inability to move (except vertical eye movement).

Damage to structure back brain stem (B) causes coma.

A technology for self-awareness?

Why create a technology for self-awaress?

1) A Grand Challenge for humanity (on the order of human flight, travel to the moon or mapping the genome).
2) Pure scientific (curiosity driven) research.
3) (personal view)A potentially powerful enabling technology for machine learning and development.

For 60 years the quest for self-Aware machines has been side-tracked in a quest for machine “intelligence”.

What do we mean by Intelligent?

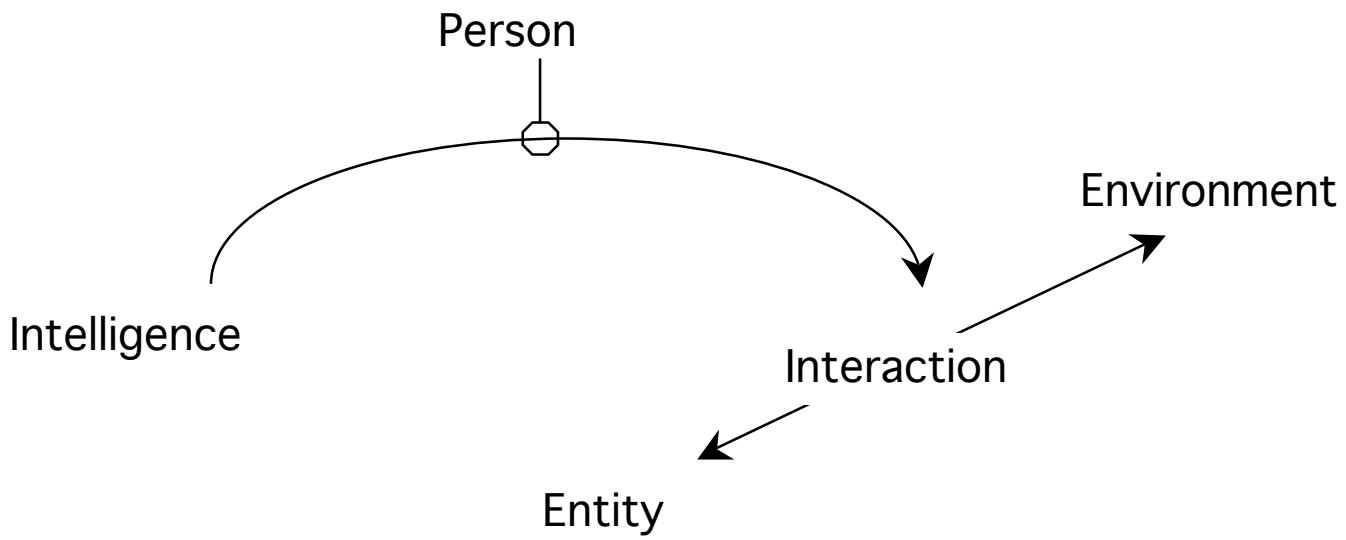
Intelligence describes the interaction of an entity with its environment.*

Intelligence is a description (an ascribed property)

Intelligence describes an entity that interacts.

*Cognitive Systems Research Roadmap (2002), European Commission, ECVision Network (David Vernon, Editor).

What do we mean by Intelligent?

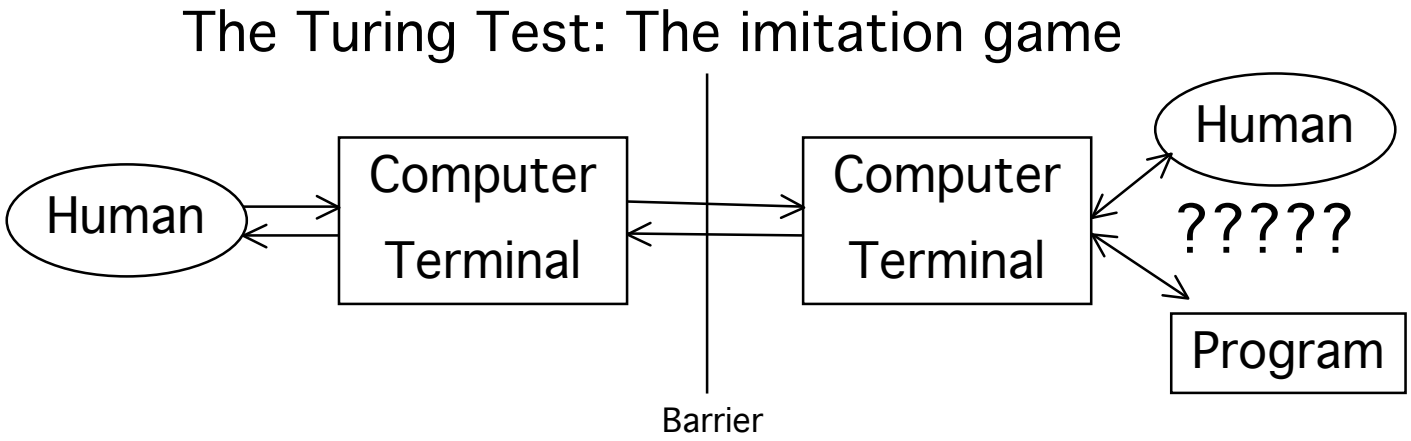


Intelligence describes the interaction of an entity with its environment.

Human level intelligence requires an abilitiy for social interaction. (the Turing Test).

Early Computing:

Turing's View of Intelligence



Alan Turing changed the question from "Can machines think?" to "Can a machine behave as a human?"

Turing claimed that a machine would exhibit intelligence if it exhibited behaviour that could not be distinguished from a person.

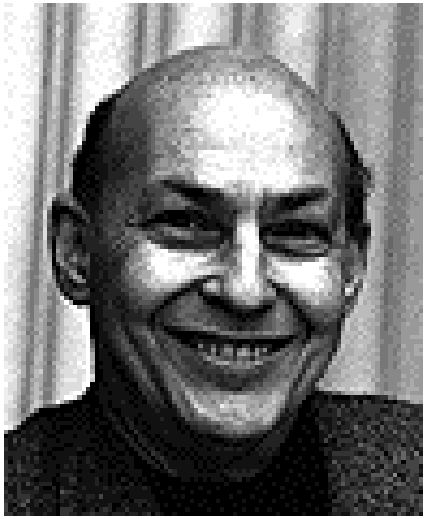
The Turing test measures abilities for social interaction.

Social Interaction for Personal Robots

The is no technology for personal robots to learn polite interaction.

- My Roomba cleaning robot is aware but not self aware.
- My Aibo robot interacts but does not have a second order model of his own sensori-motor maps.
- There is no technology to endow a robot with a "sense" of "self"

Computer Science: Marvin Minsky

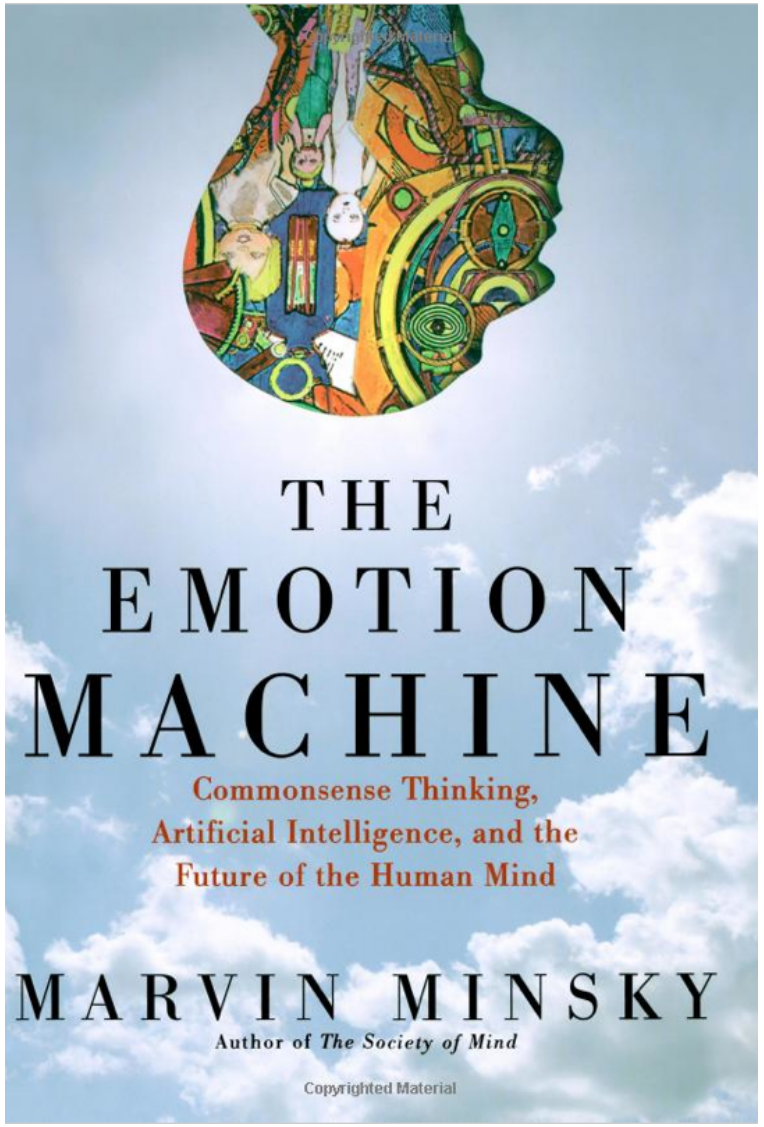


Marvin Minsky

Professor MIT

PhD in mathematics Princeton (1954)

Fundmental contributions to AI, cognitive psychology, mathematics, computational linguistics, robotics, and optics.



Social Common Sense

Minsky distinguishes

Common sense: The collection of shared concepts and ideas that are accepted as correct by a community of people.

Social Common Sense: shared rules for polite, social interaction that govern behavior within a group

Situated Social Common Sense: Social common sense appropriate to the current situation.

Situated Social Common Sense

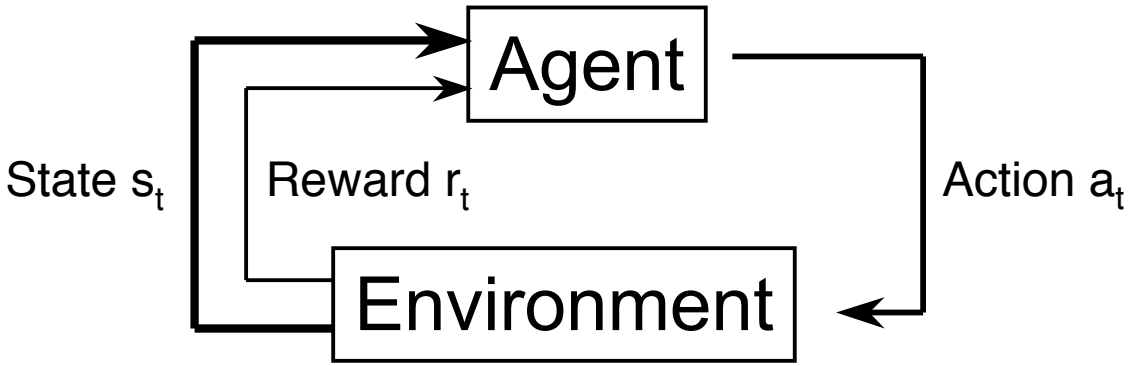
Assertions:

- Politeness is a problem of Situated Social Common Sense
- Politeness requires understanding social situation

Social Common Sense varies over individuals and groups.

⇒ Social Common Sense must be learned through interaction

Reinforcement Learning



Algorithm to learn a policy π: *situation* → *action*

Q-Learning [Watkins 89] :

- Learns policy and Q-Value at the same time
- Based on temporal different and learning rate.

Fundamental Hard Problem: Credit Assignment; Assigning rewards to actions.

Q-Learning

Estimates the value of state-action pairs

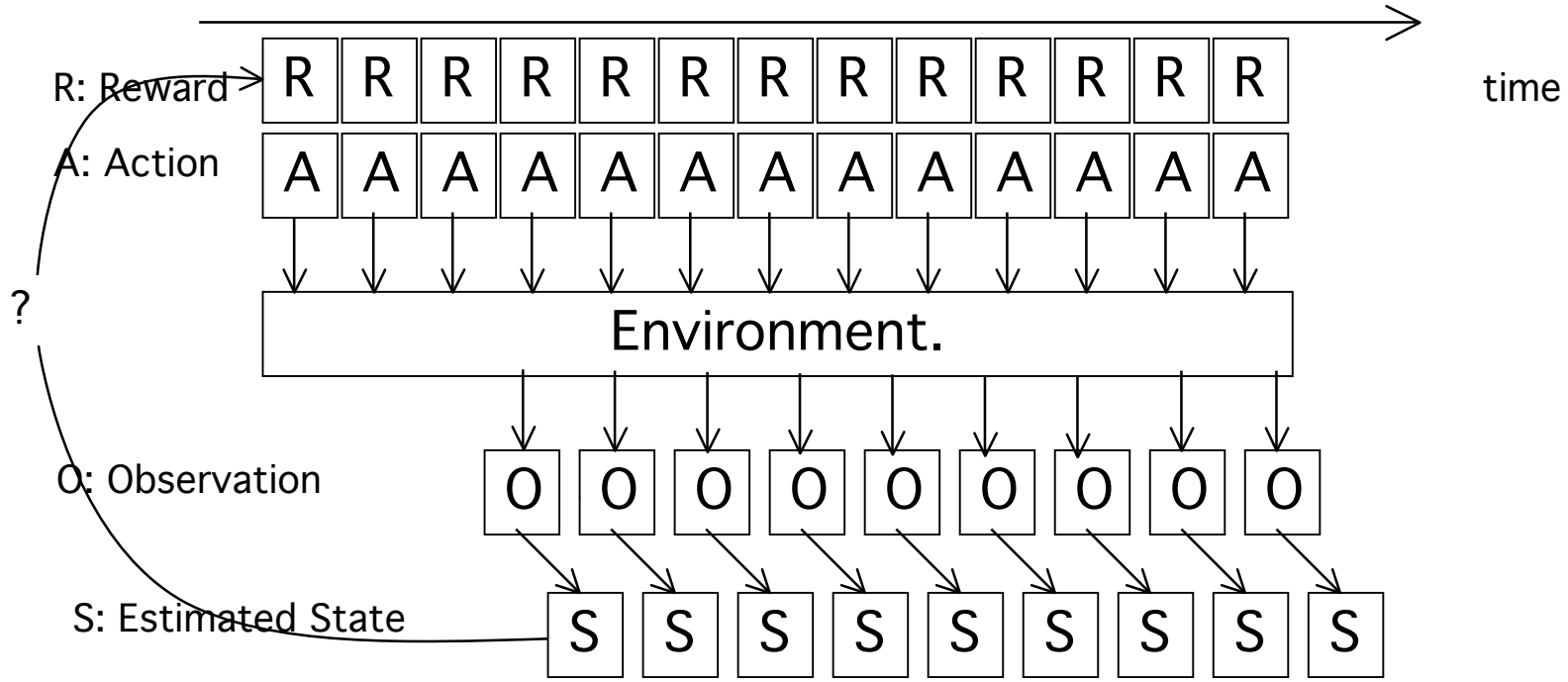
$$Q_{t+1}(s_t, a_t) = (1 - \alpha_t) \cdot Q_t(s_t, a_t) + \alpha_t \left(R(s_t, a_t) + \gamma \max_{a'} Q_t(s_{t+1}, a') \right)$$

Asynchronous Q-learning: Variable learning rate with (s, t)

$$\alpha_t = \frac{1}{(1 + t_{s,a})^w} \text{ and } w \in \left(\frac{1}{2}, 1 \right)$$

Immediate reward fails completely in complex interactions.

The Credit Assignment Problem



Reinforcement learning works fine for simple actions with fixed delay effects.

Reinforcement learning is impractical for social interaciton because of the complexity of the actions and the results.

Why is Social Learning hard?

Learning rate should depend on many social factors, including

- Social Context
- Time of Day
- Emotional stimulation
- Motivation, Attention
- Nature of reward
- Surprise

Hypothesis (personal): The autobiographical Self" may facilitate learning social common sense by providing abstraction.

Combined with case-based learning, self-awareness may permit learning for social interaction.

Why is Social Learning hard?

Fundamental Problem

The Knowledge Barrier:

The extreme complexity of human activity and individual preferences makes credit assignment for idividual actions impossible

Proposed Solution

Use autobiographical self is a method for organising abstract models of behaviour for learning from social interaction.

Possible Impacts on Technology

Self awareness can be used to provide:

- Enhanced system reliability
- Machine Learning for social interaction
- Learning for complex behaviours.

Possible Application Domains

- Personal Robots (Self-aware appliances)
- Teaching machines
- Self-aware Smart Homes
- Self-aware Intelligent Vehicles
- Self-aware Smart Cities

Impacts on Society

- Artificial systems with beyond-human intelligence.
- Social interaction with self-aware robots and services an as part of Human Development. (teaching machines training children?)
- Extension of “human rights” to Machines.
- Legal responsibilities for autonomous self-aware machines?

Conclusions

1. The invention of a technology for autonomous self-aware machines is a long grand challenge for humanity
2. Scientific investigation in this area is “pre-paradigm”
3. Emergence of a working theory is (finally) within our reach.
4. A working theory requires convergences of ideas and methods from a variety of disciplines.
5. Success may have long term impact on humanity and the human condition.