From Feedback to Consciousness

Ricardo Sanz

Autonomous Systems Laboratory
Universidad Politécnica de Madrid

Machine consciousness. Complexity aspects
Exystence Topical Workshop
Torino, September 29 – October 1, 2003

Abstract

- Control engineering is a technical activity where the concept of feedback plays a central role.
- Basic controllers perceive the reality in a simple way and determine control actions based on deviations from desired states.
- The search for autonomous behaviour goes beyond a simple schema, requiring complex control structures to deal with complex world situations.
- Modern controllers have performance requirements that are difficult to meet because the controller should deal with sensor multimodality, environment uncertainty, faults, plant complexity, etc. Artificial intelligence plays a central role and control engineering can now be seen as mind engineering.
- Next steps in complex control engineering are strongly related with system self-reflection, deep understanding of situations and the emergence of selves.
- In simple terms, complex controllers are becoming conscious.

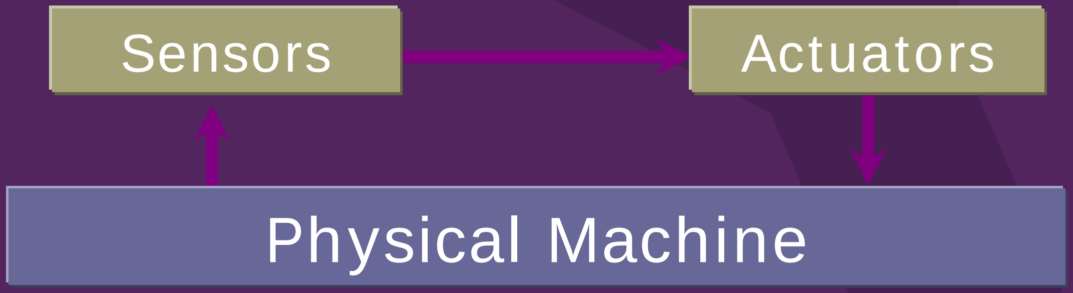
Contents

- Complexity raising
- Integrated Reflective Controllers
- A Theory of Consciousness

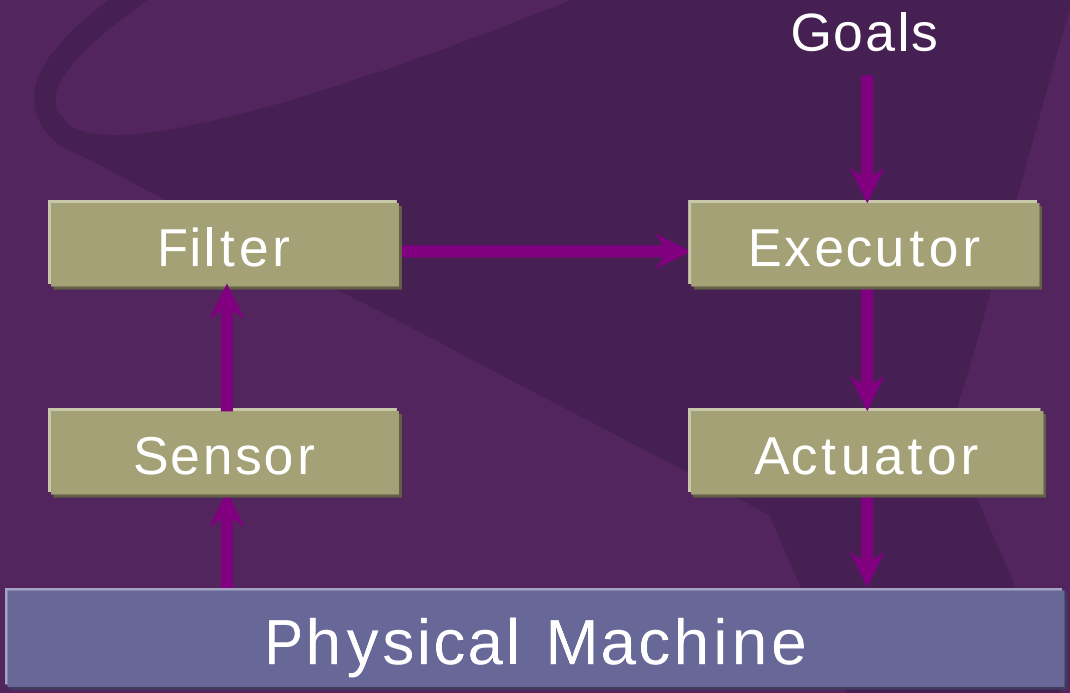
Complexity raising

Summarising trends in controller architectures

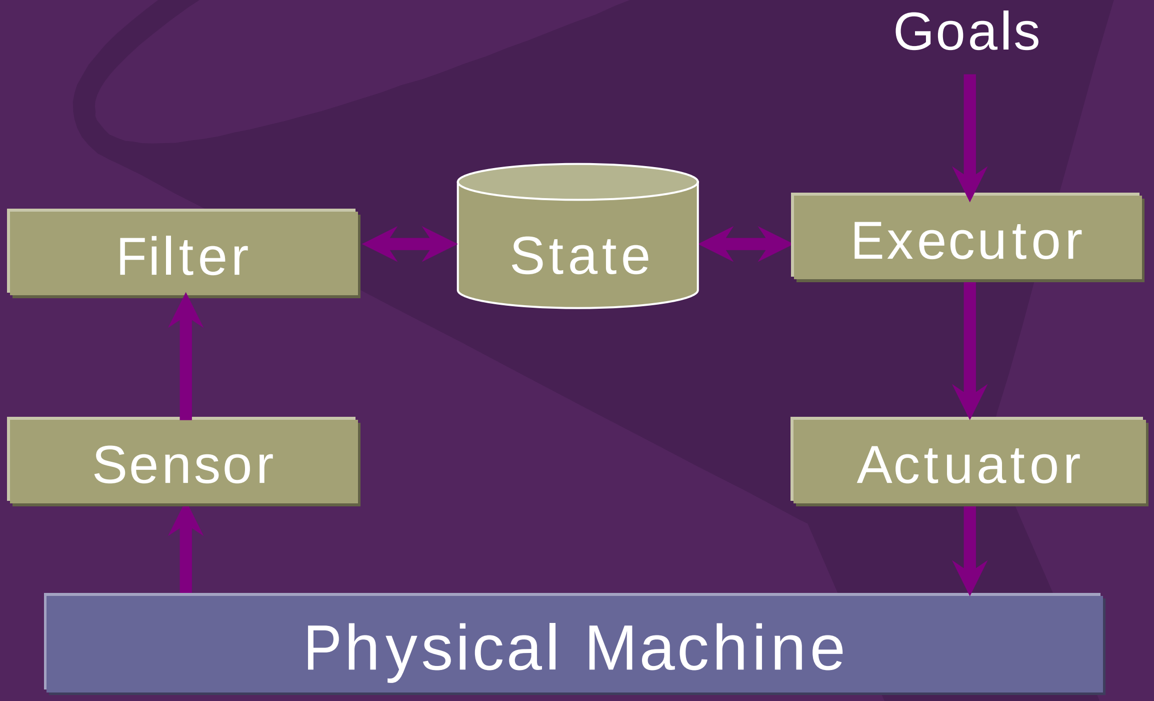
Simple feedback



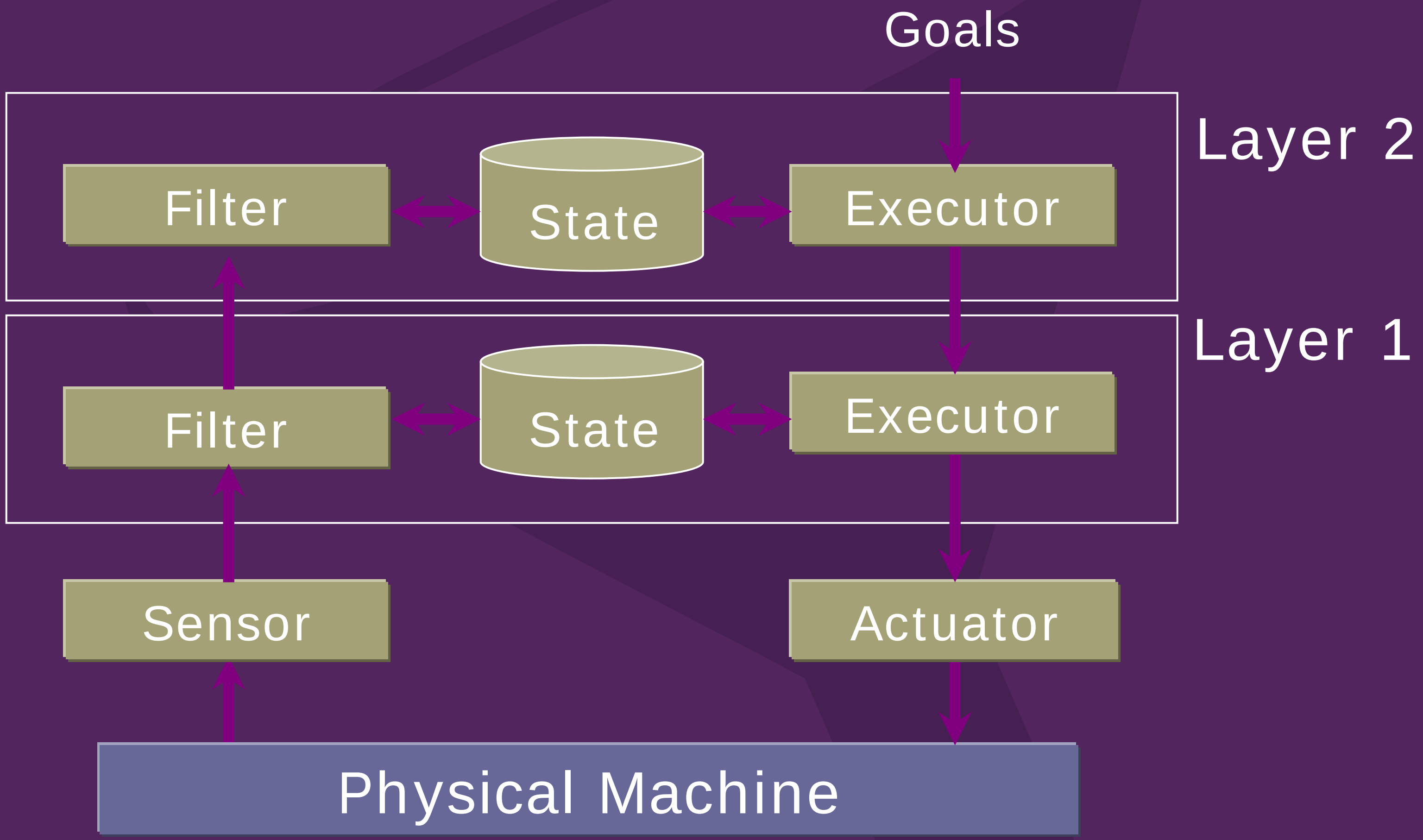
Enhanced feedback



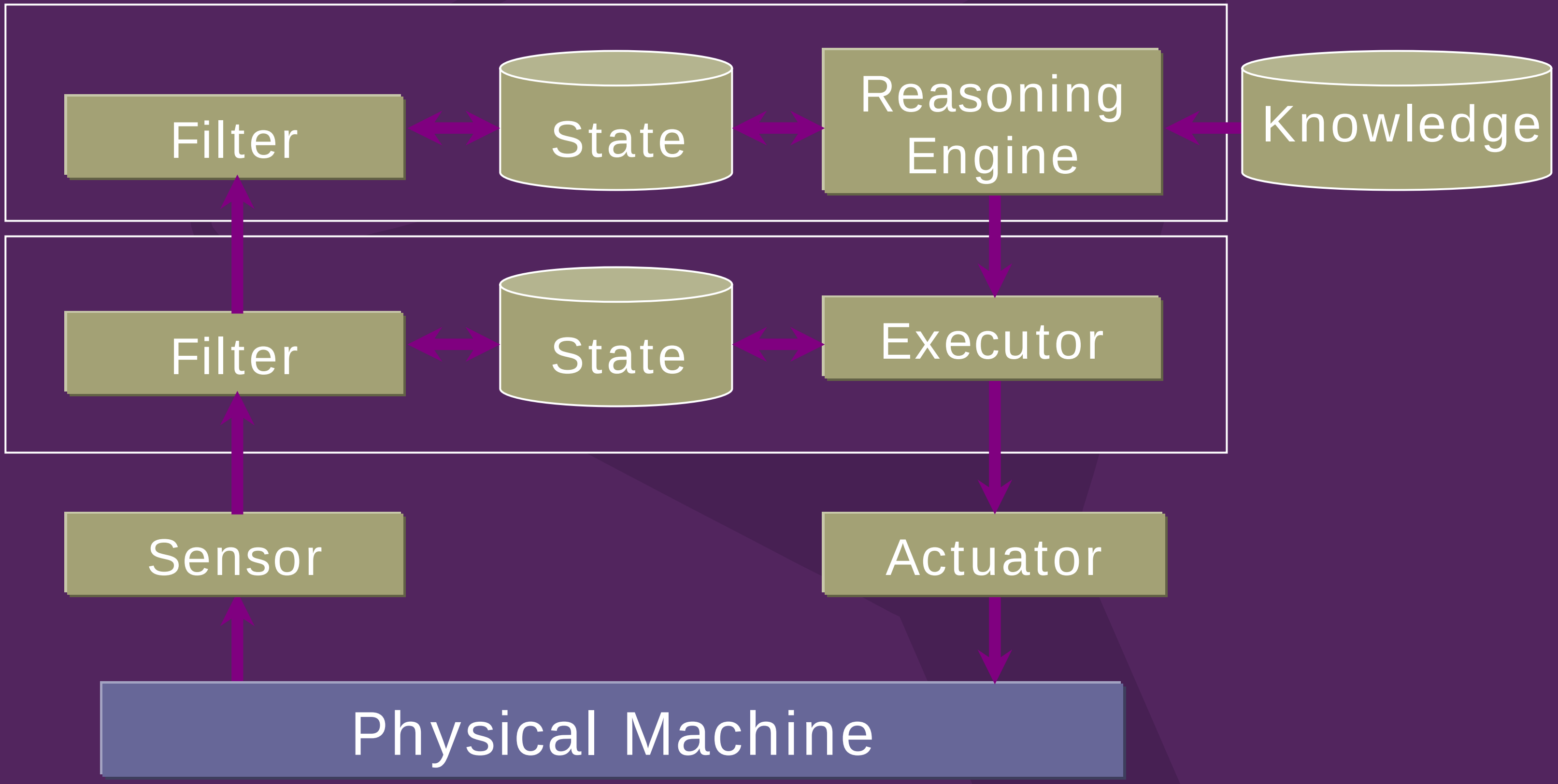
Stateless vs Stateful



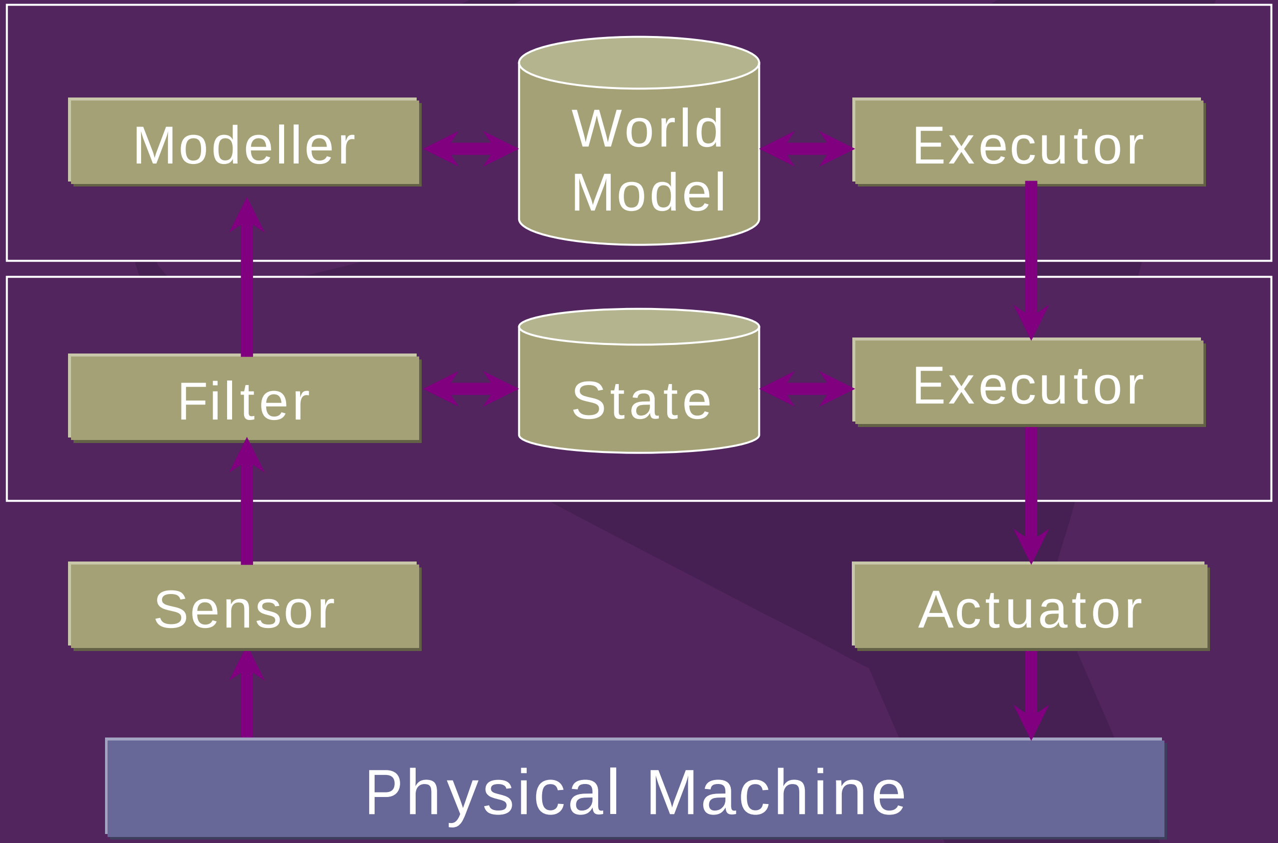
Layering


Goals
Layer 2
Filter
State
Executor
Layer 1
Filter
State
Executor
Sensor
Actuator
Physical Machine

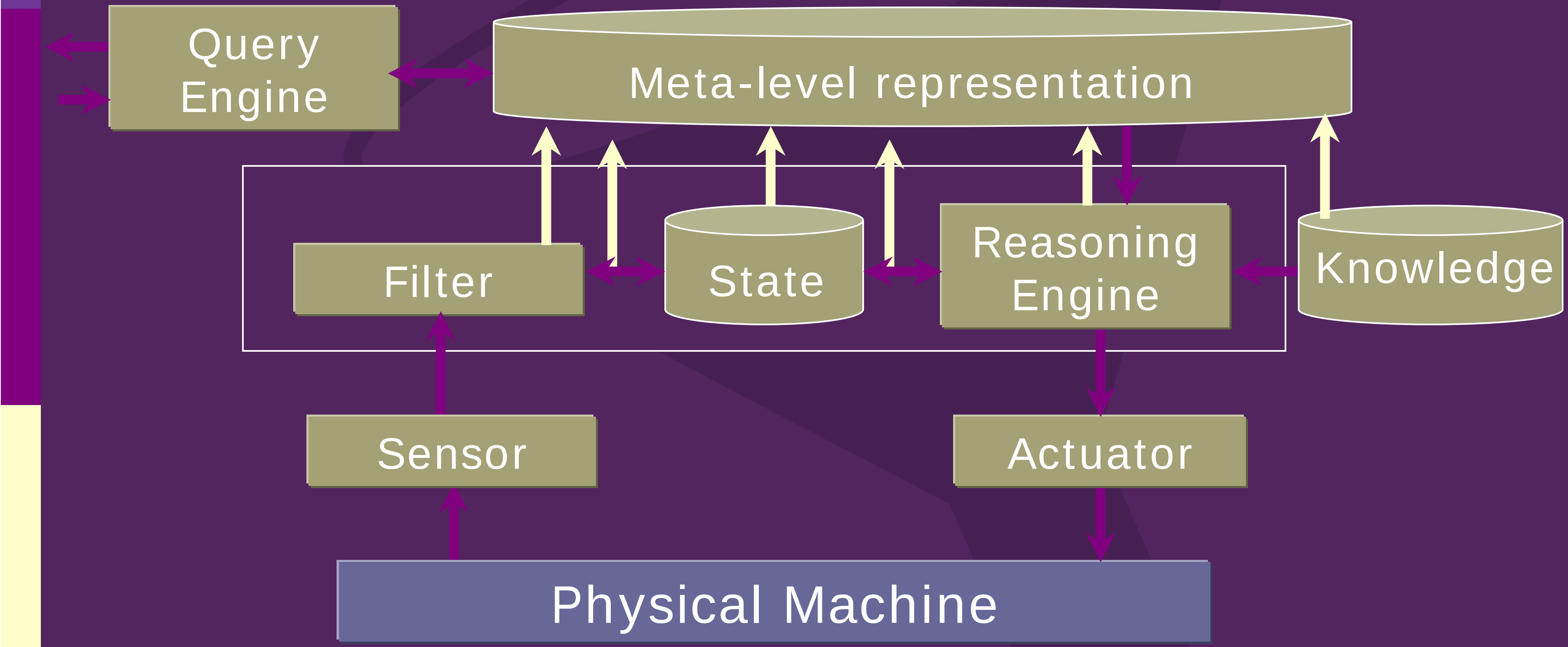
Deliverative/Reactive



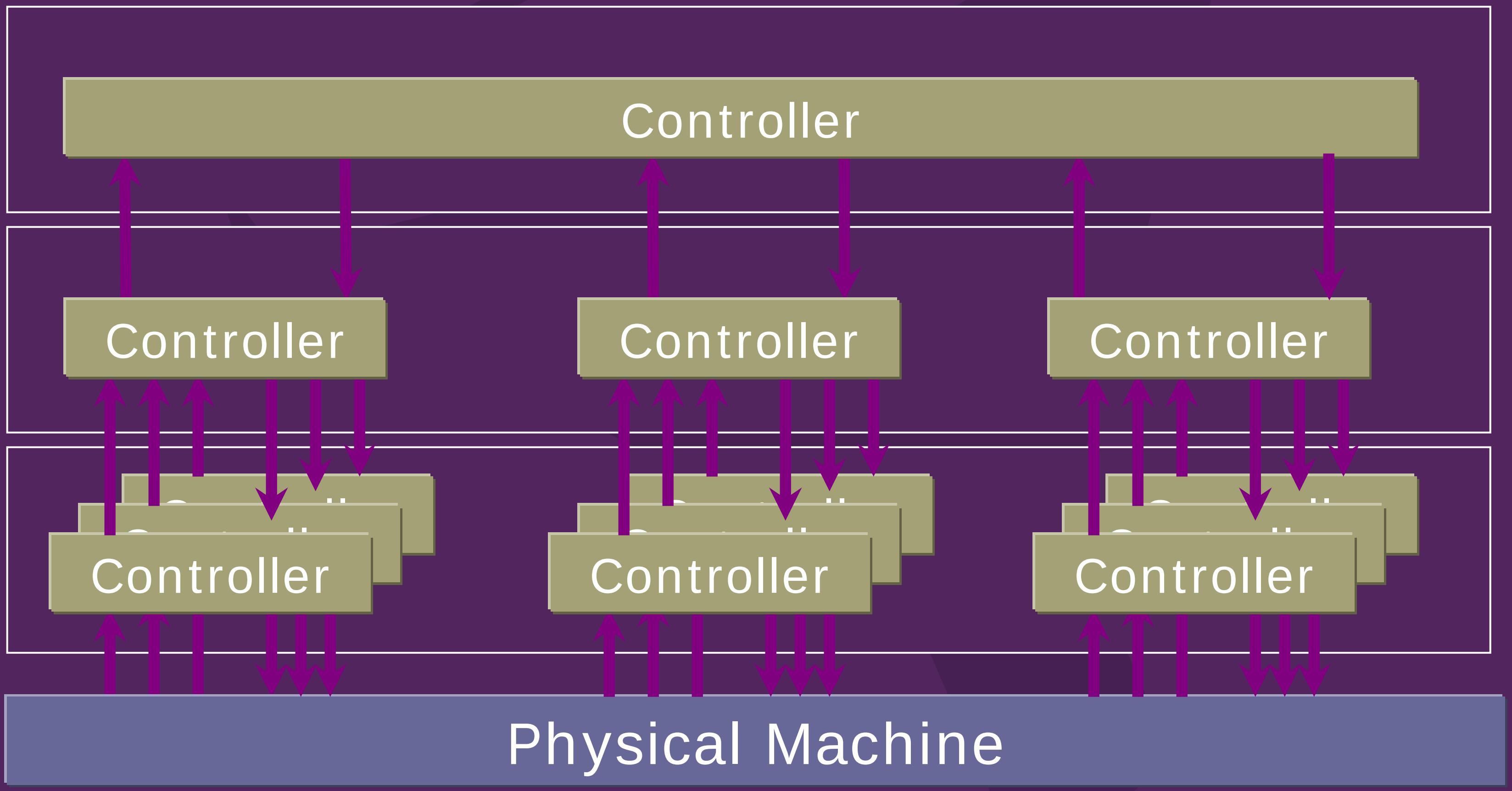
Model-based control



Introspection and Reflection



Hierarchy and heterarchy



Beyond “Normal” Agents

- Dependable control agents do have requirements well beyond what is considered “normal” intelligent function:
  - Real-time behavior
  - Embeddability
  - Evolvability
  - Ugradeability
  - Robustness

Robust design

- The suggestion that robust design is the primary source of complexity is motivated by the observation that for most biological and technological systems, the vast majority of components are present for robustness rather than for basic functionality of the organism or machine.

  [Reynolds et al. 2001]

Two Notions of Complexity

- “Complexity emerges in systems that are otherwise internally homogeneous and simple”.
  - Self-organized criticality (SOC) and the edge of chaos suggests that large-scale structure arises naturally and at no apparent cost through collective fluctuations in systems with generic interactions between individual agents.

- “Complexity is associated with intricately designed or highly evolved systems”.
  - Highly optimized tolerance (HOT) emphasizes the role of robustness to uncertainties in the environment as a driving force towards increasing complexity in biological evolution and engineering design.

Integrated Reflective Controllers

Systems that reason about themselves

Mechanisms for robustness

- H_2, H_∞ (robust control)
  - The system tolerates small displacement from design conditions
- Redundancy
  - Increase robustness up to a limit where the increase in dependability is less than the new induced risks
- Fault-tolerance
  - Copes with plant changes due to faults
- Reflection

Autonomic Systems

(as IBM sees them)

- Adapts to changes in its environment
- Strives to improve its performance
- Heals when it is damaged
- Defends itself against attackers
- Exchanges resources with unfamiliar systems
- Communicates through open standards
- Anticipates users' actions

- Possesses a sense of self

SciAm May 06, 2002

Next steps in complex control

- System introspection and reflection
- Deep understanding of situations
- Self-healing beyond adaptation
- Emergence of integrated selves

Multiresolutional reflective control

- Revonsuo: Biological systems have “multiple levels of organisation, forming a hierarchical, causal mechanical network”
- Industrial controller evolution is mimicking biological mind evolution
- Now, we’re in the phase of creating conscious controllers (even when most control engineers don’t know or don’t say in public)

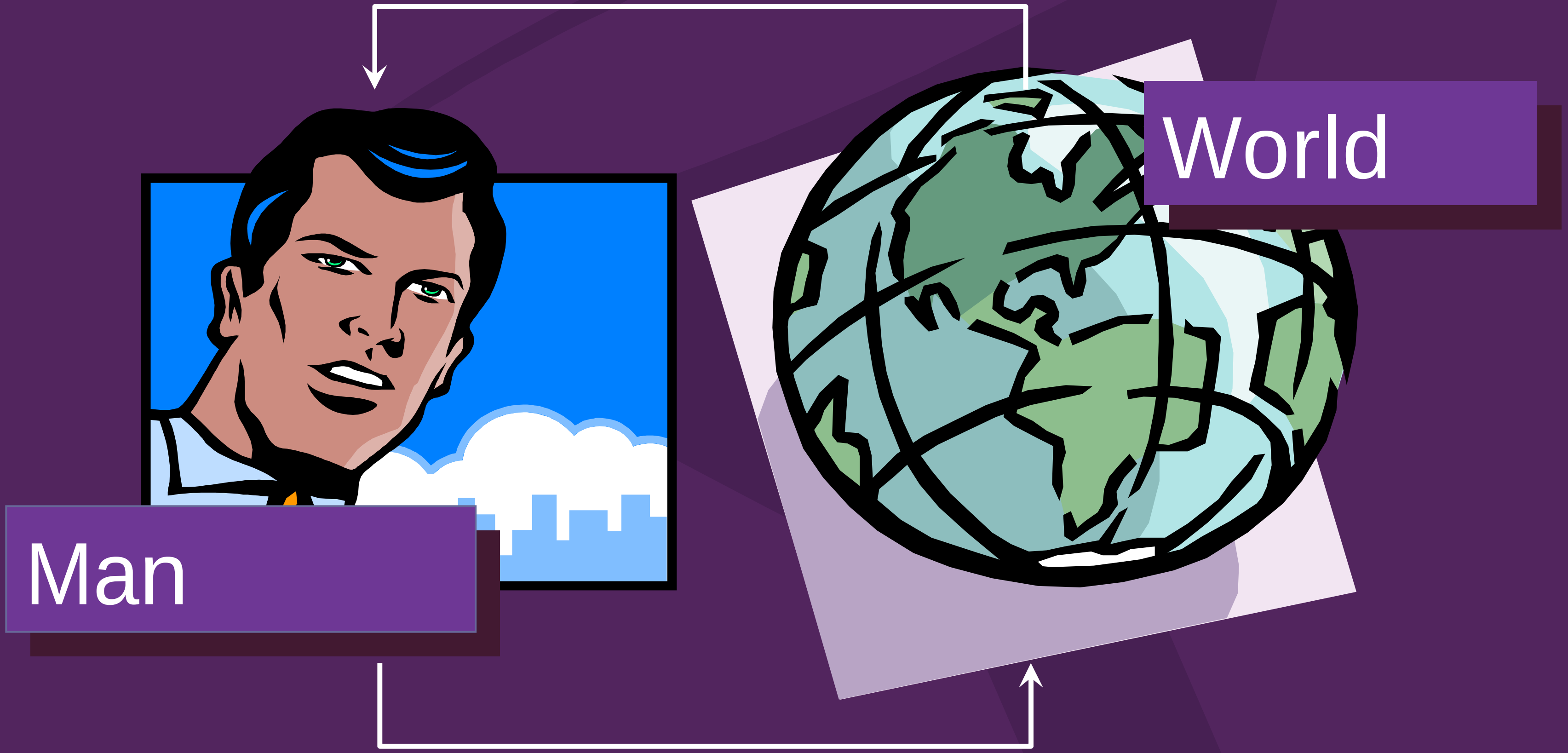
A Theory of Consciousness

Based on control designs

First Assessment

- There is a single emerging model of consciousness
- Neuroscientific/psychological data corroborates this model
- Varying visions are just views of this core model coloured of personal backgrounds

Man on His World



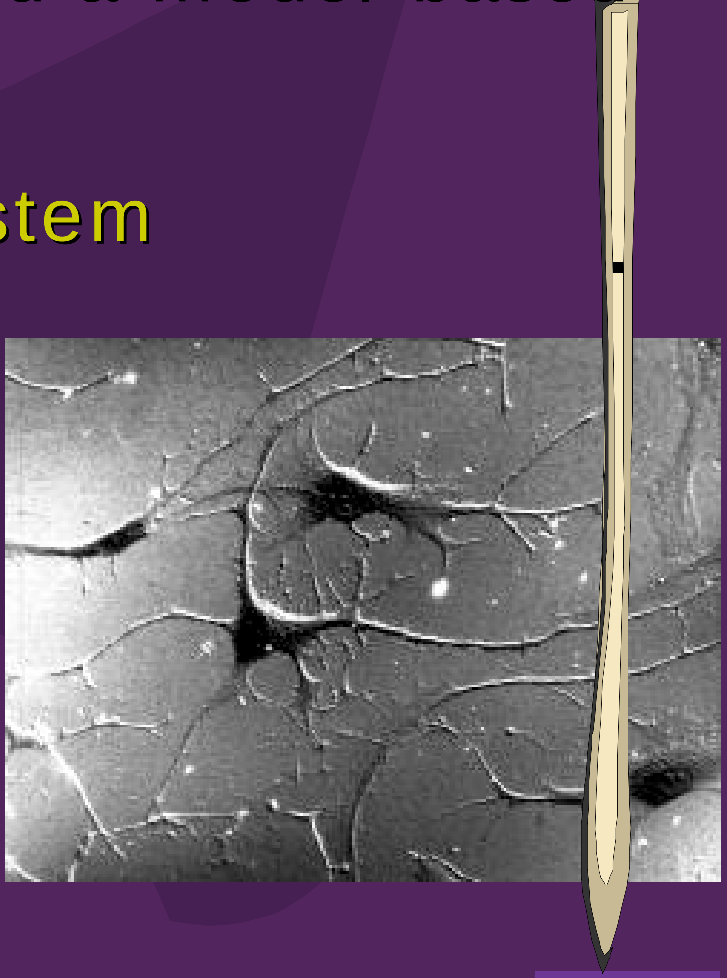
Man

World

The modeling machine

- Evolution has engineered a model-based learning controller:

  the Central Nervous System



- This machine generates different types of models to properly act in the world

Naïve models of reality

- Judging that an animal will not mind being killed if it is not offended, Eskimos take various ritual precautions before, during, and after the hunt.

- The rationale (the behavioral model of the world+agent) lies in the belief that animal spirits exist independent of bodies and are reborn: an offended animal will later lead his companions away so that the hunter may starve.

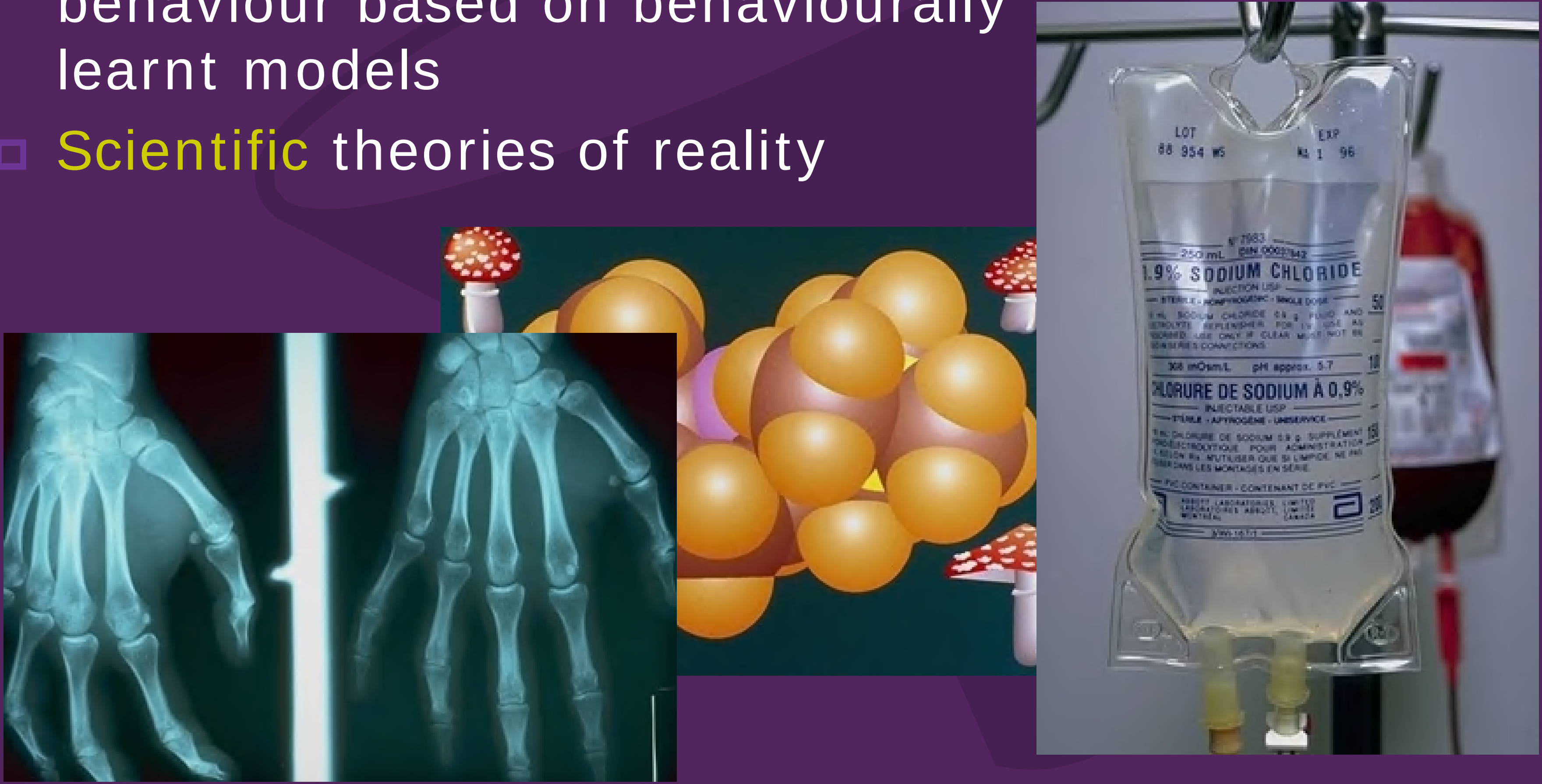
- Just projections of what is best known: the self

Survivor Models

- Elementary, ad-hoc, experience-based causal models of reality
- Examples: agriculture, mating, Micronesian navigation (rowing to move the islands to certain positions in the horizon)
- Cleermans: "representational systems that can be adaptively modified by ongoing experience"

Deep models

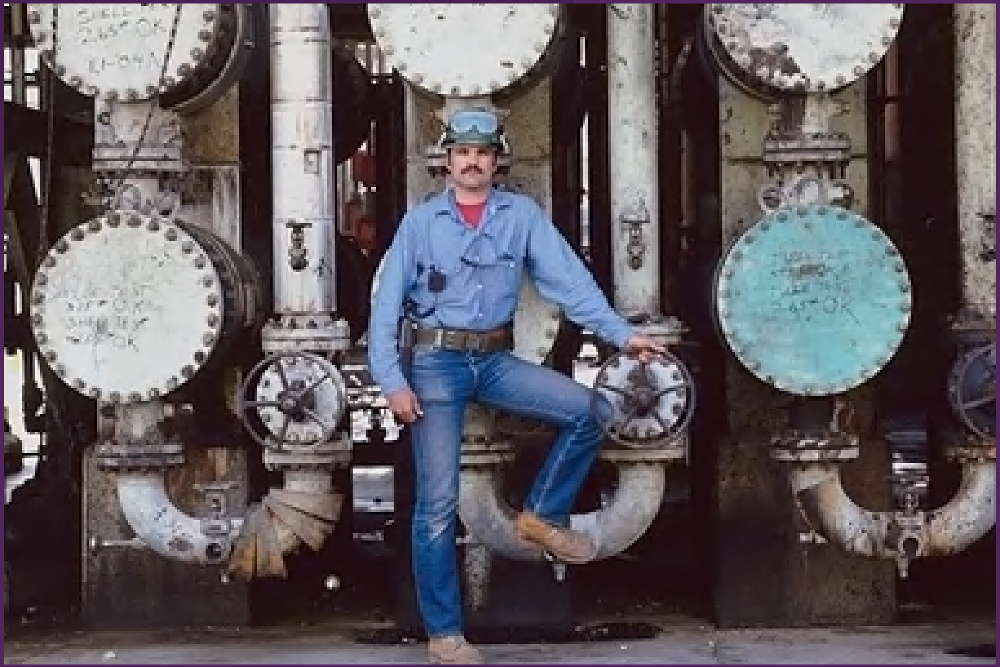
- Behaviour based on deep models outperforms behaviour based on behaviourally learnt models
- Scientific theories of reality



The machine of the world

- Science and technology have established themselves as the best models and tools to control the machinery of the world

- Wigner: "*The unreasonable effectiveness of mathematics in the natural sciences*"



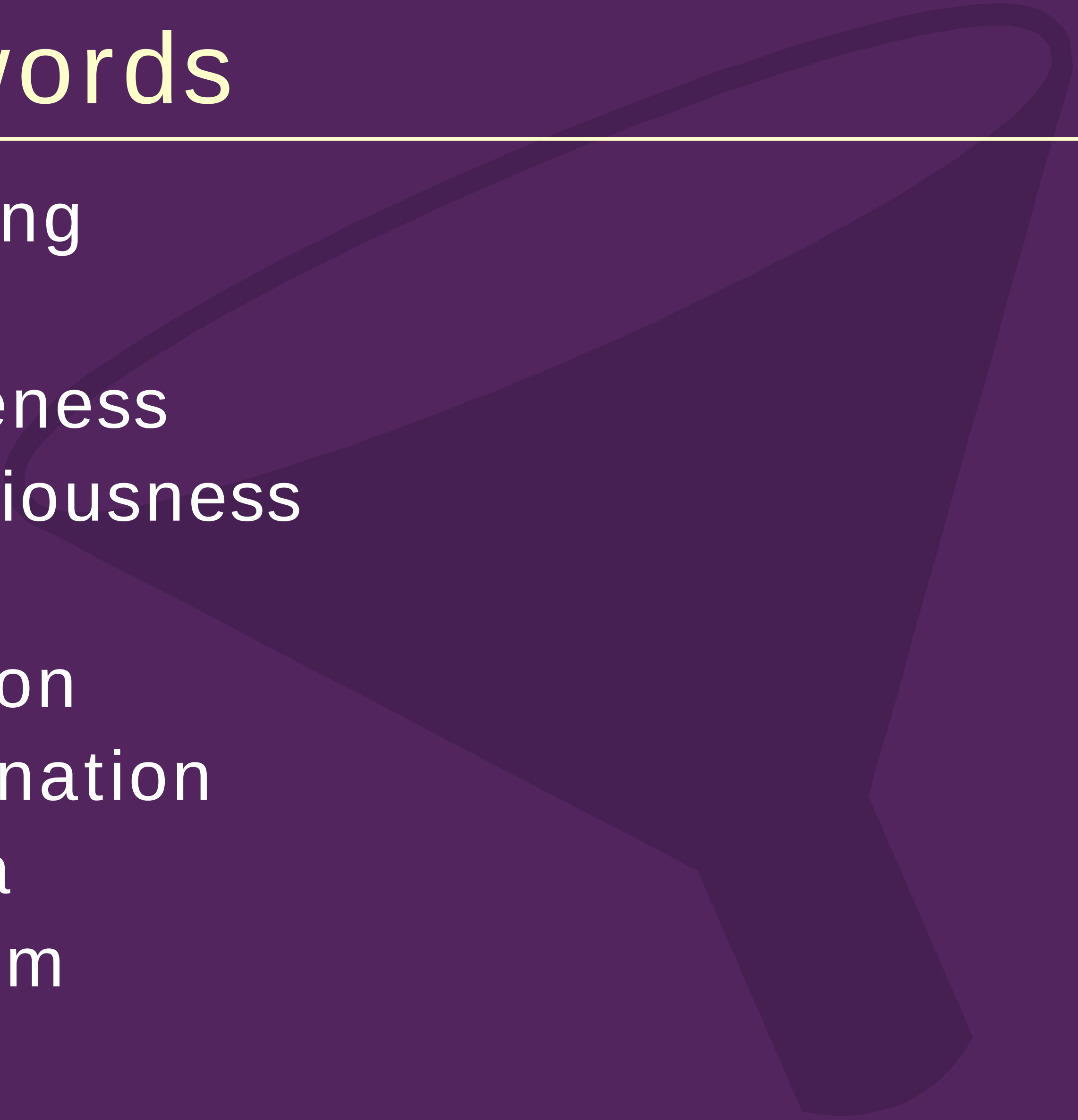
Central design

- The mind is a multiresolutional adaptive-predictive model-based control system that has reflection properties

!!!

Hot words

- □ Meaning
- □ Value
- □ Awareness
- □ Consciousness
- □ Self
- □ Emotion
- □ Imagination
- □ Qualia
- □ Wisdom

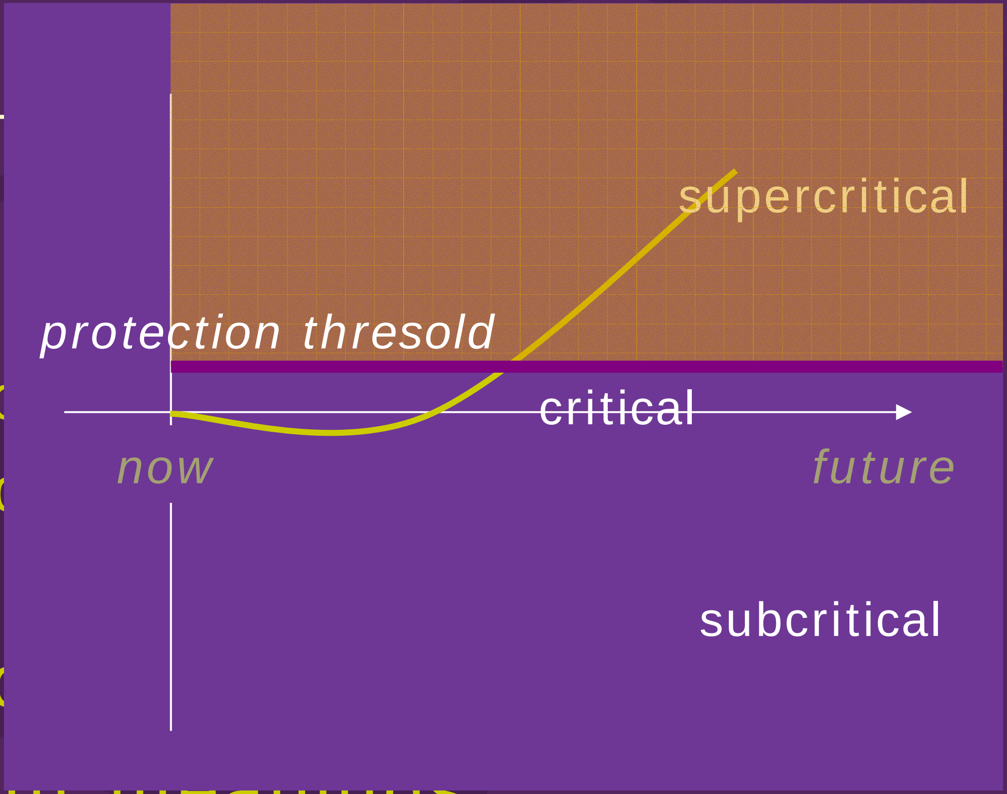


Meaning

- Autonomous systems:
  - Generate meanings from data (typically from sensory inputs)
  - Use their continuously updated mental models to control behaviour
- Meanings are equivalence classes of agent + relevant world trajectories in state-space in relation with agent's value system

Example 1

- Consider:
  - a nuclear reactor
  - the primary contro
  - the primary protec
- What's the meani measure of neutro
- The're two different meanings
  - for the control system
  - for the protection system



Example 2

- Consider that you're driving along a road going to Rome
- Consider the meaning of a road sign saying "Rome to the right"
- If you become aware of the sign, the value of your future along the road changes completely

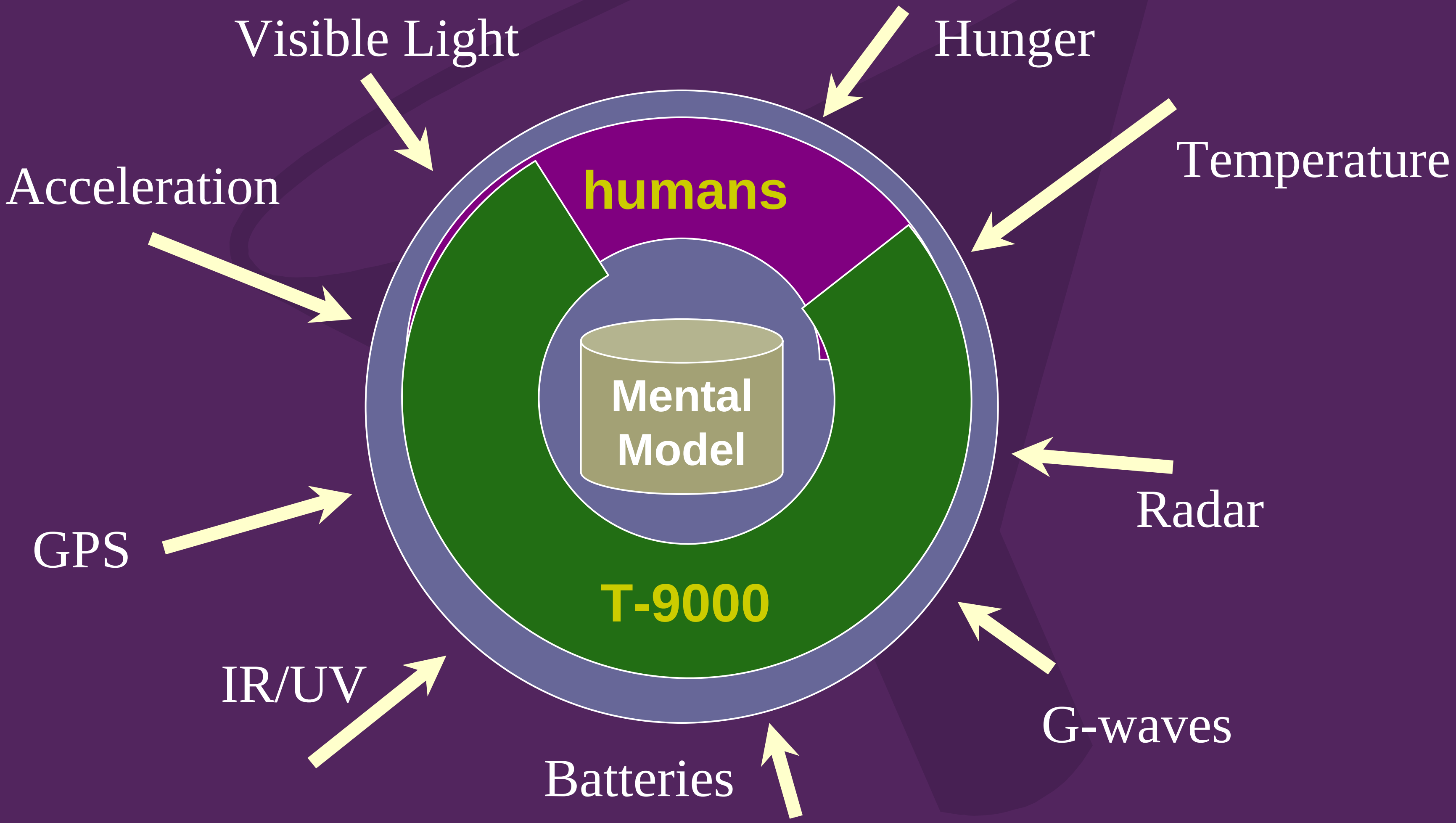
Value

- Values are computed of states (in the present or in the future) and used to generate emotions
- Prediction engines are critical for generation of potential futures

Awareness

- A system is aware if it is generating meanings from perceptions (including proprioception)
- Perception updates the inner models
- Bear in mind that awareness/meaning is not an static thing but full of dynamical content due to the dynamical nature of the models

Awareness spectra



Consciousness

- A system is conscious if “I am aware” is valid in the present state of affairs (it is generated from the perceptual flow, i.e. the system is aware of itself).
- Lacombe: “It is impossible to separate awareness, consciousness and understanding”.

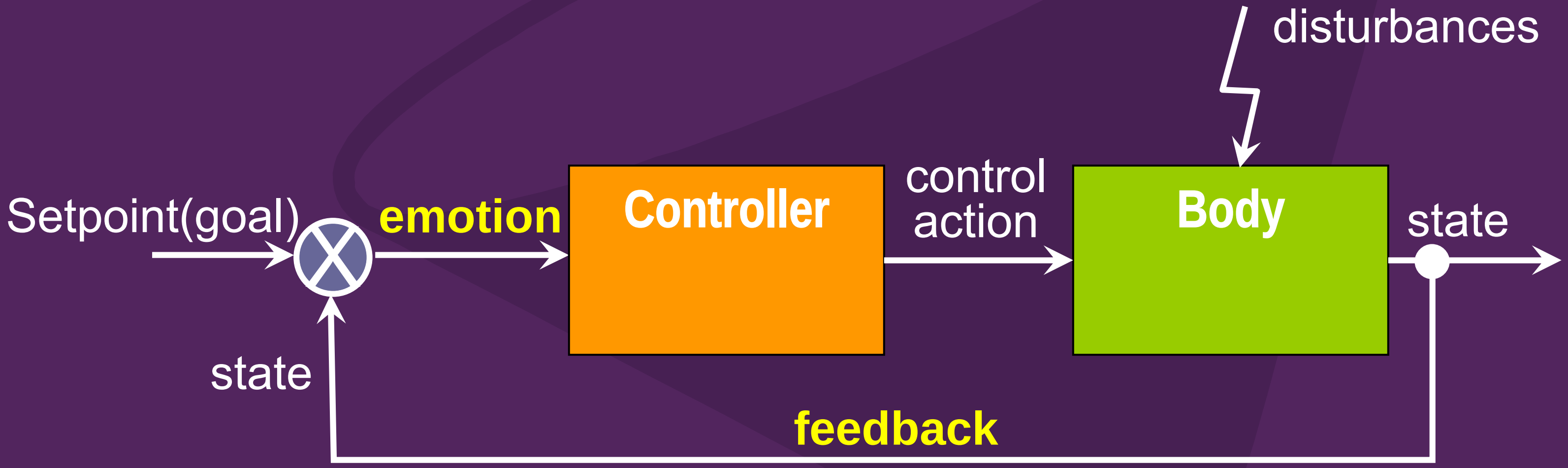
Self

- When the sub-models of the being are integrated and encapsulated the self emerges as a simple compound fact about a particular object

- Obviously, this fact has all the meaning generation capabilities of any other fact

Emotion

- Emotions = Inputs to controllers



- Error in feedback controller
- Downward action in layered controllers

Emotion

- Emotions are hierarchical as are the control topologies
- Emotions can have different representation mechanisms (including implicit representations)
- Emotions drive action

Imagination

- Projections into the future including counterfactuals
- It uses the internal dynamical models
- The mechanism for exploring value space of potential futures

Qualia

- Some half-baked ideas but without a sufficiently clear description yet (need to put more grey matter on it)
- (well, it is the hard problem, uh?)

Wisdom

- Aware controllers are able to generate meanings for their bearers
- Wise controllers can generate meanings for others
- Do not underestimate the difficulty of this task for a learning controller

Strong points of this model

- Unifiable
- Machine applicable
- Explains other related phenomena: e.g. attention
  - Of what do you calculate potential effects when resources are scarce?
  - Only of those pieces that most assuredly can affect your future: focus of attention
  - Holland: “simulate only the part of the world that can mostly affect the agent”

Another Strong Point

- This model provides a metric
- It is possible to calculate the degree of coverage of future trajectories
- It makes possible the comparison of awareness levels of systems that are in the same conditions (i.e. experiencing the same sensor space, including inner space)

Summary

- Mind is a multiresolutional phenomenon of model-based adaptive-predictive control
- Minds generate and use dynamic models
- At any resolution level, meaning generation generates awareness
- At any resolution level, mind reflection generates consciousness
- We -usually- only can talk about the upper level in ourselves

Thanks

Questions ?