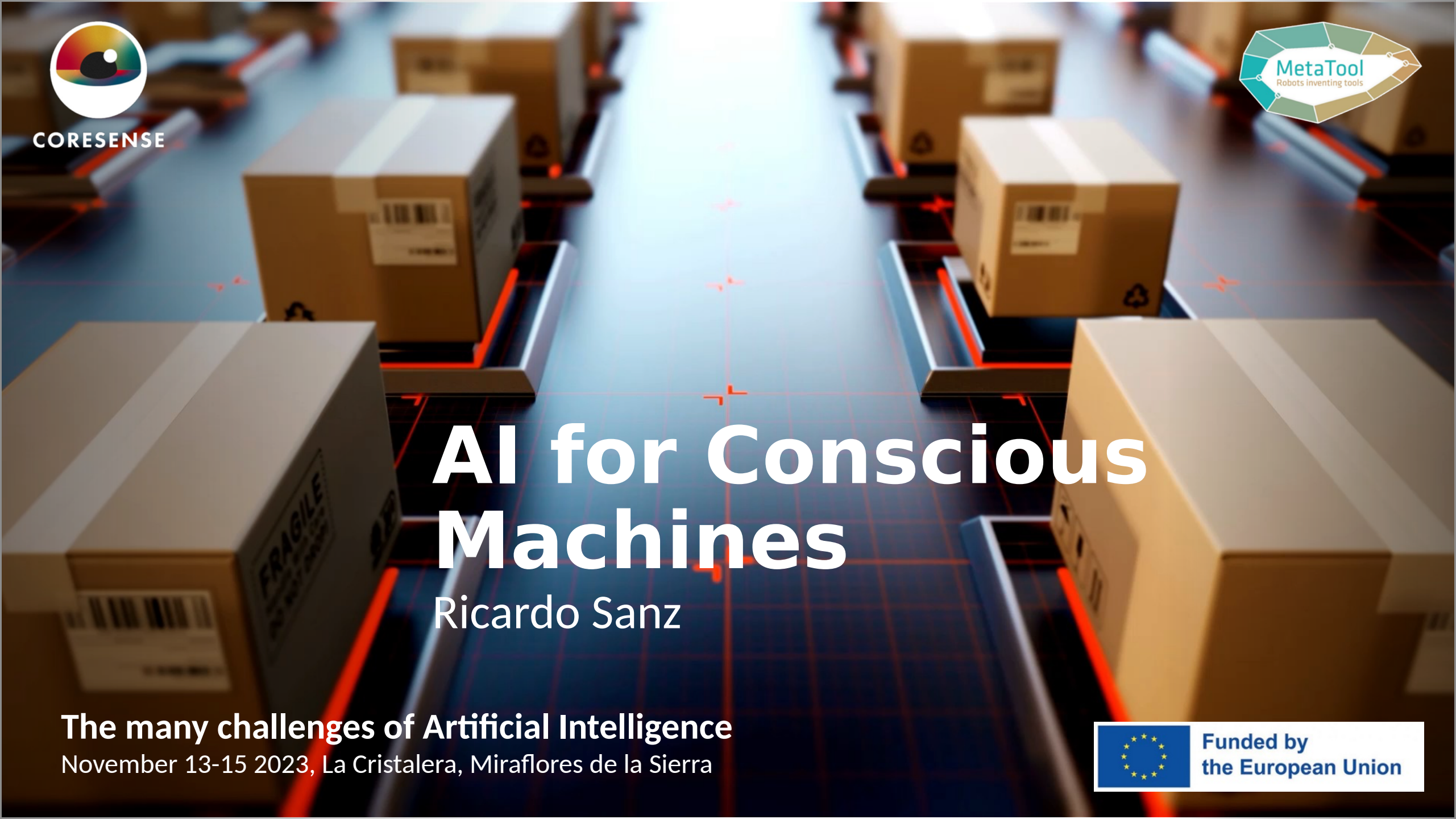


AI for Conscious Machines

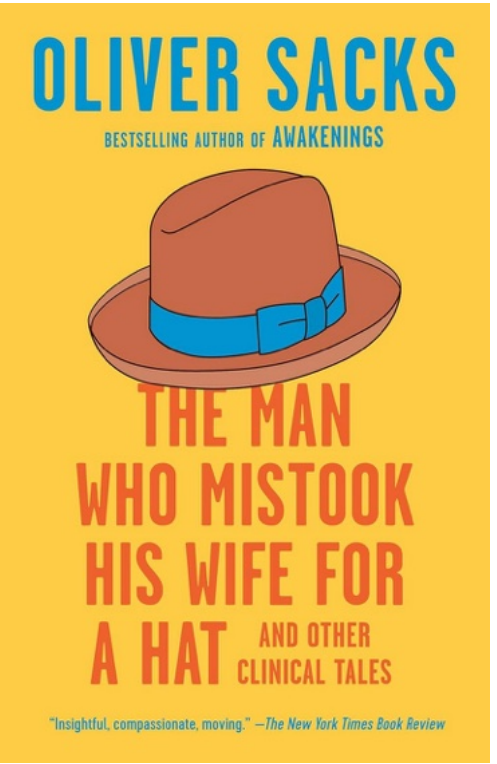
Ricardo Sanz

The many challenges of Artificial Intelligence

November 13-15 2023, La Cristalera, Miraflores de la Sierra

Funded by the European Union

AI in the real world: Tesla autonomy



See also:

[The 737 MAX MCAS]

What is the “essence” of the problem?

most of what we build is fake intelligence

most of what we build is **shallow** intelligence

LLMs don’t have knowledge, but a statistical summary of knowledege

AGI

Two Challenges for AI

UNDERSTANDING

AWARENESS

Content

- About me and the theme
- Themes on robot consciousness
- Two projects

Ricardo Sanz

UPM Autonomous Systems Laboratory

autonomy –
giving oneself the behavior laws

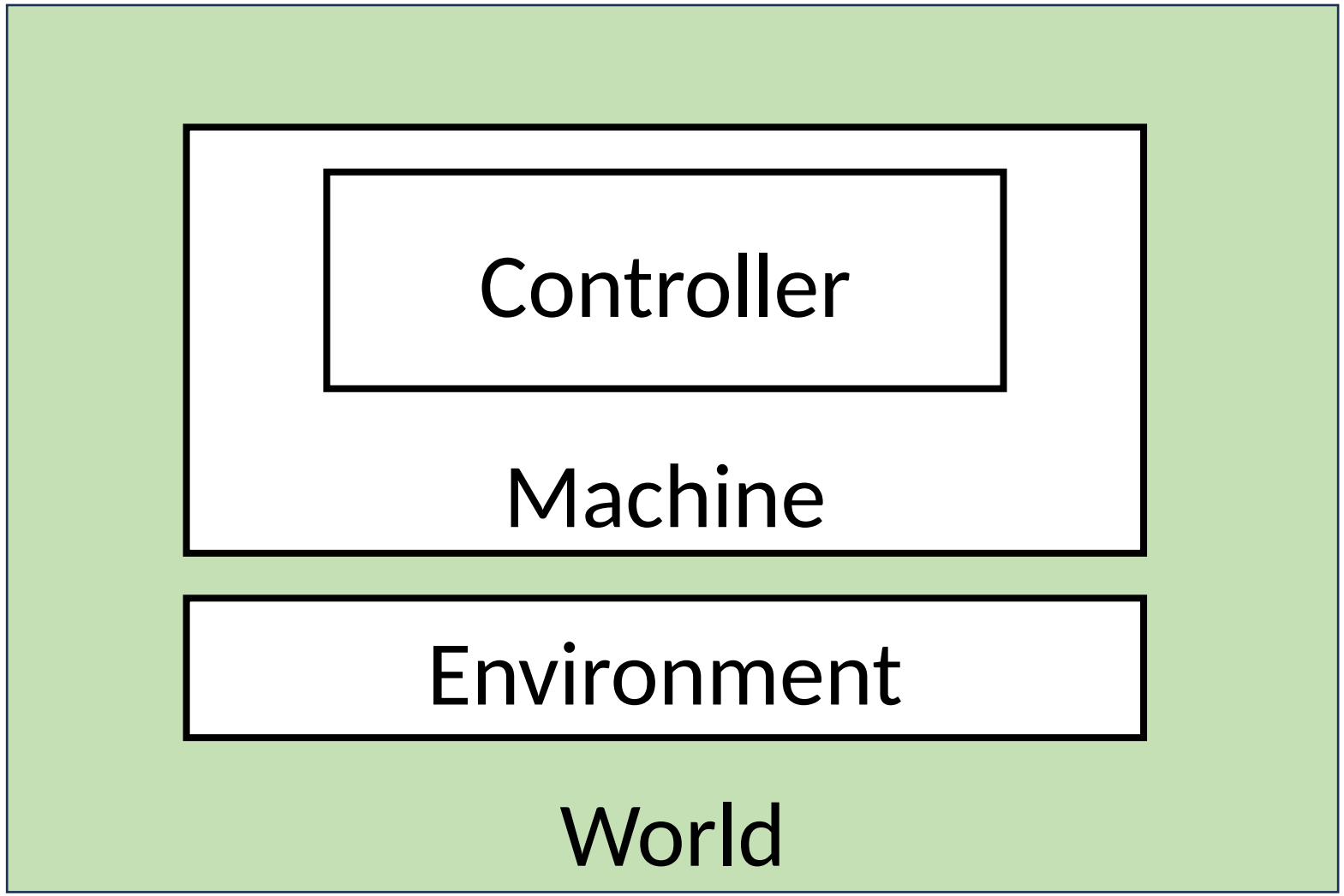
from the Greek autos (self) and nomos (law)

Autonomy for what?

- Cost reduction, human safety and improved **performance** were the main factors behind the drive for improved autonomy in the past.
- We seek **mission autonomy**.
- During the last years, however, a new force is gaining momentum: the need for **augmented dependability** of complex systems.
- We seek **robust autonomy**.
- Machines are not / should not be autonomous in a very *strictu senso*. They must be **under control** in all situations.
- We seek **bounded autonomy.**

A machine

Control engineering




A380
Continental
GM
vector
ibeo

My research focus

- Move the responsibility - mission, robustness, resilience, safety – **into the system itself**

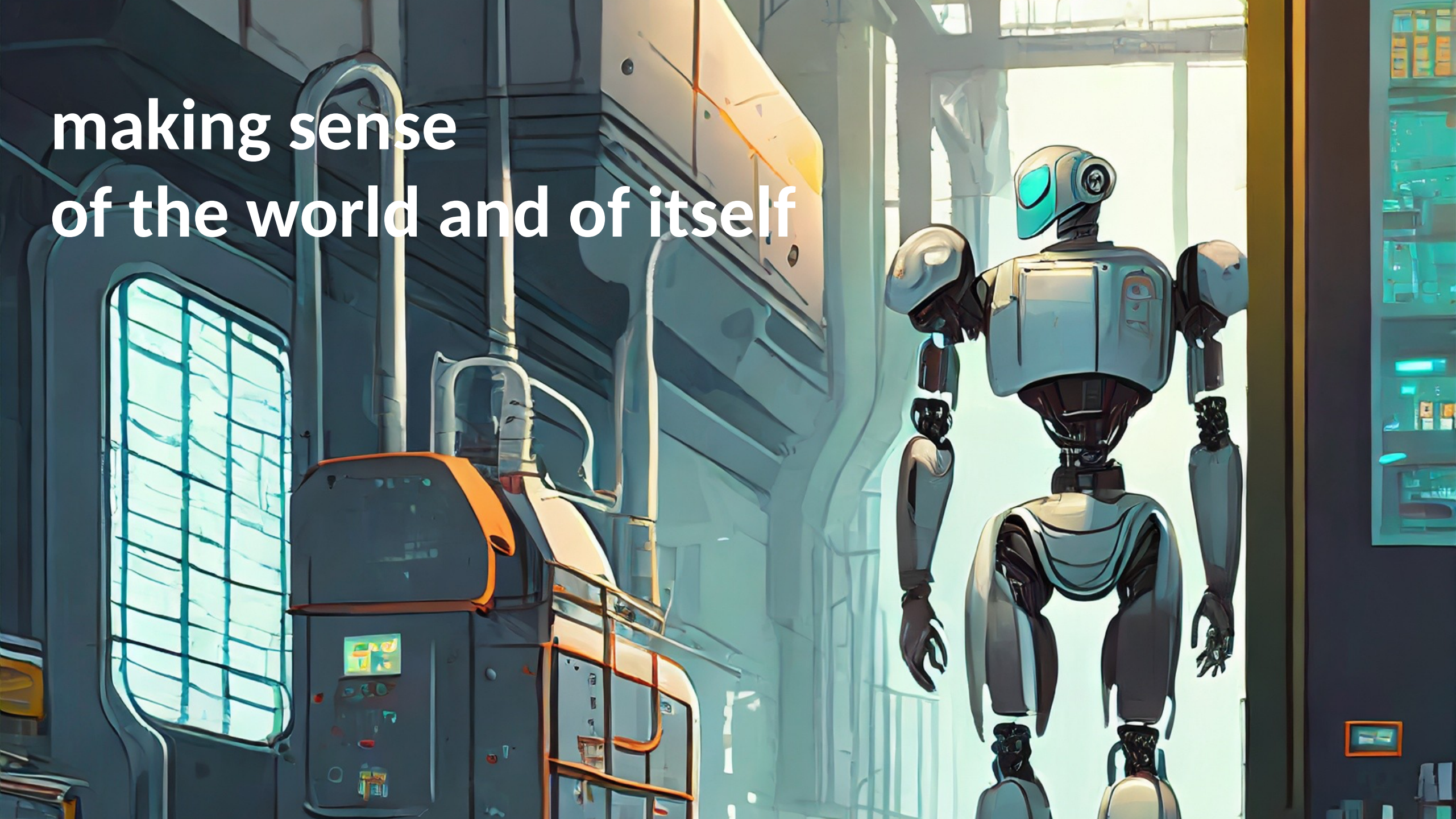
Self-aware AI

Skynet !!!

Self-X as Research Target

- We aim at computer-based controllers for systems which are world-aware, and self-aware, self-configuring, self-optimising, self-healing, self-protecting, and self-adapting (**self-x**).

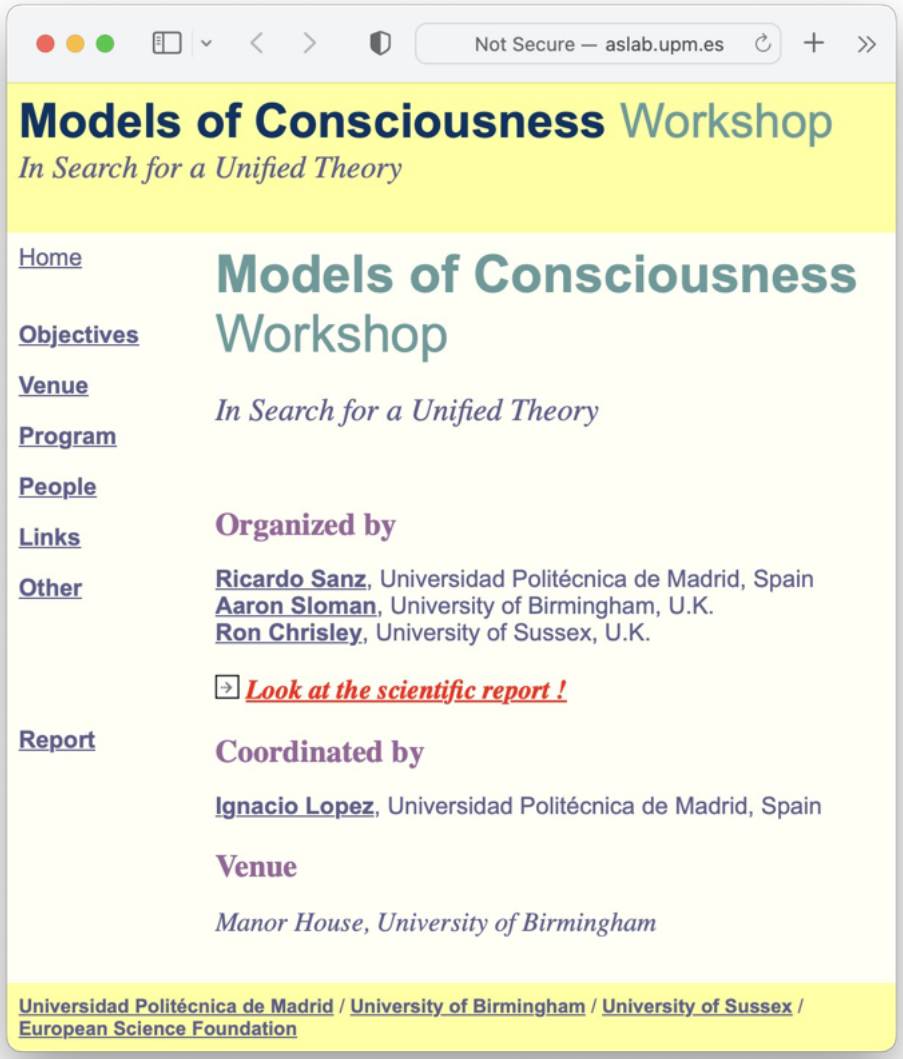
- We seek an increase of functionality, constructability and resilience in many system functions by means of the incorporation into the very controller of mechanisms for having '**deep understanding**' and '**self-awareness**'.


making sense
of the world and of itself

- **Twenty years ago (2003)**, with money from the **European Science Foundation** we (R.Sanz, A.Sloman, R.Chrisley) organised a focused workshop on consciouness at Birmingham:
- Models of Consciousness Workshop

  http://www.aslab.upm.es/events/MOC2003/
- We had 28 researchers from seven European countries and **19 talks** by consciousness-relevant researchers (both in bio and AI):

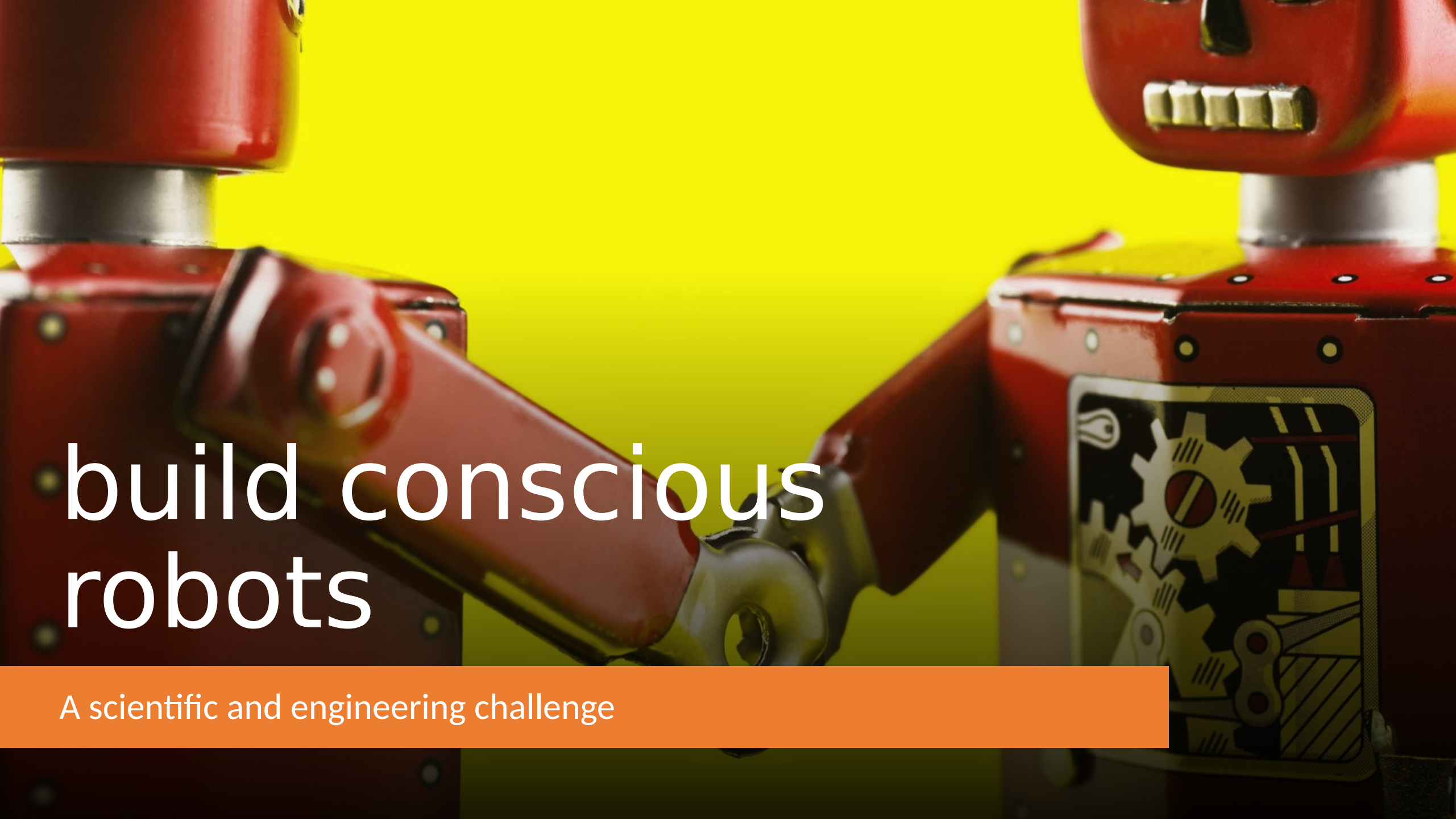
  François Anceau, Axel Cleermans, Jim Doran, William Edmonson, Petros Gelepithis, Pentti Haikonen, Germund Hesslow, Owen Holland, Jacques Lacombe, Riccardo Manzotti, Peter Redgrave, Geraint Rees, Antti Revonsuo, Miguel, Salichs, Ricardo Sanz, Murray Shanahan, Aaron Sloman, John Taylor, Tom Ziemke.



EUROPEAN SCIENCE FOUNDATION

- The Models of Consciousness workshop tried to advance the elaboration of a **unified scientific theory of consciousness**. A core topic of the workshop was unified theories of natural and artificial consciousness and this workshop focused on the particular aspects of models of consciousness that are also suitable for implementation, i.e. **theories of consciousness that can support the construction of conscious machines** and also serve as explanation of the experimental data about the natural kind of consciousness.

- Now: More researchers; money; projects; neuroscience; robots.
- However, after 20 years, **the theory has not advanced substantially**.
- Neither the implementation.
- Especially in the domain of robot consciousness.



build conscious robots

A scientific and engineering challenge

Maybe Turing was right

- “There are, however, special remarks to be made about many of the disabilities that have been mentioned. The inability to **enjoy strawberries and cream** may have struck the reader as frivolous. Possibly a machine might be made to enjoy this delicious dish, but **any attempt to make one do so would be idiotic**.” [Turing-1950, p.448]

Turing, A. M. 'Computing Machinery and Intelligence', Mind, vol. 59 (1950), pp. 433-460.

Maybe Gabriel is right

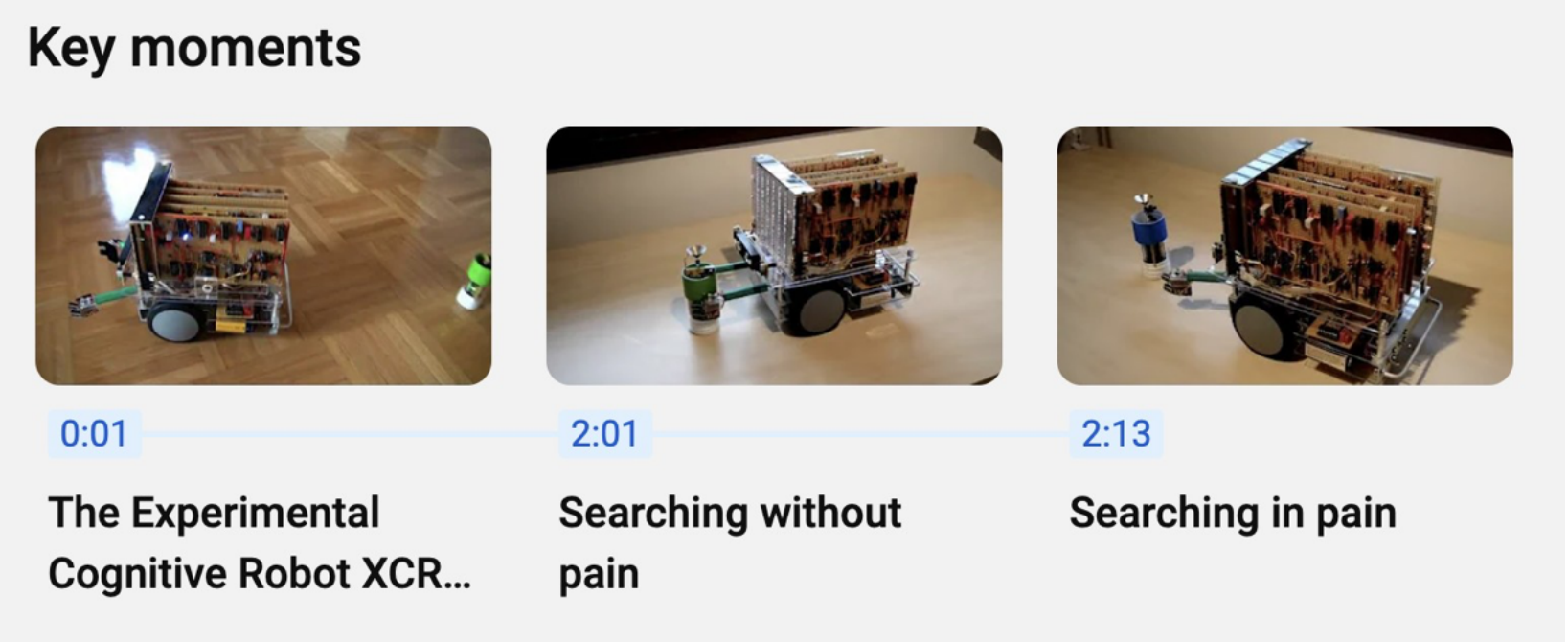
- “Could a Robot be Conscious? The shortest answer to the question posed in my title is: **No**.”

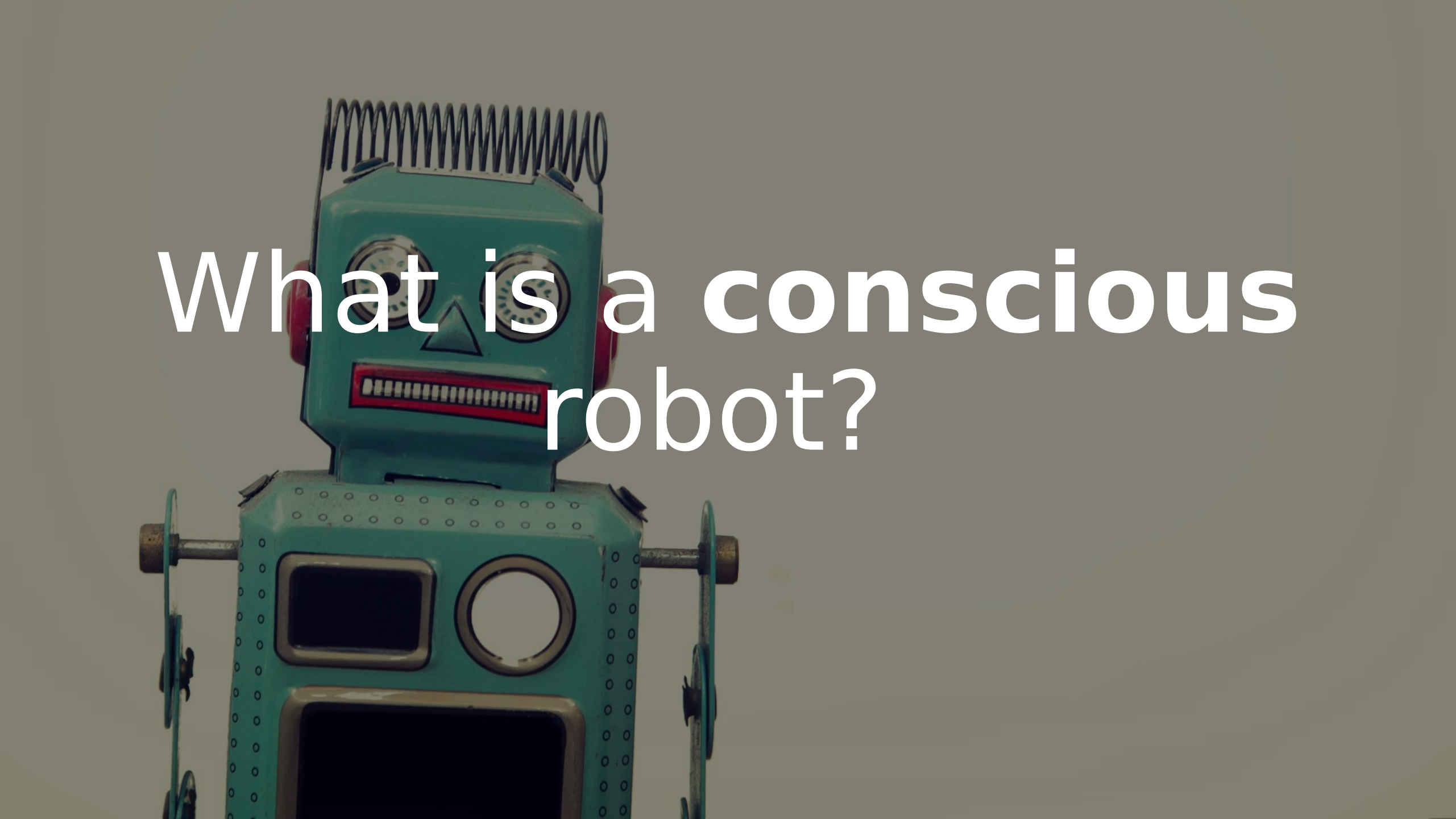
Gabriel, M. Could a Robot Be Conscious? Some Lessons from Philosophy. In Robotics, AI, and Humanity. Science, Ethics, and Policy, by J. von Braun, M.S. Archer, G.M. Reichberg, M. Sánchez (eds), (2021).

Haikonen’s XCR-1

In this video (https://www.youtube.com/watch?v=t87QXtgfChg) we can see robot XCR-1 from my friend Haikonen, **searching “in” pain**.

In pain?




What is a **conscious** robot?

Aleksander & Dunmall: Axioms

PERCEPTION: I am in the middle of an “out there” world.

IMAGINATION: I can recall ‘out there’ worlds and imagine worlds.

ATTENTION: I am only conscious of that to which I attend.

PLANNING: I imagine doing future things.

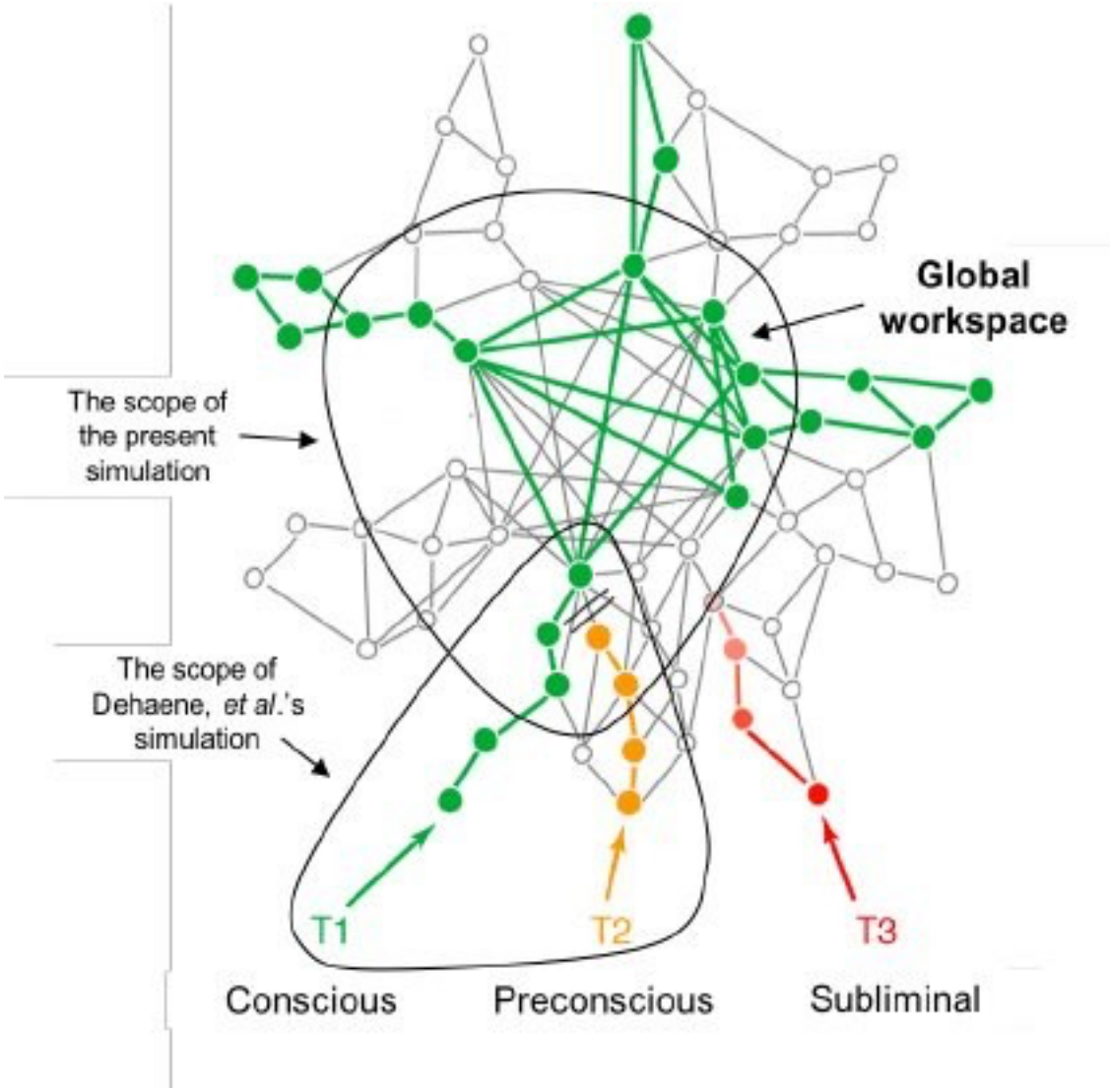
EMOTION: My emotions guide the selection of my plans.

No robot doing this yet?

Aleksander & Dunmall: Axioms and Tests for the Presence of Minimal Consciousness in Agents, Journal of Consciousness Studies June, 2003)

Shanahan GNW

A Global Neuronal Workspace model based on Baars' Global workspace Theory and Dehaene's neural model of GWT.



Shanahan, M. (2006). A cognitive architecture that combines internal simulation with a global workspace. *Consciousness and Cognition*, *15*(2), 433–449.

Dehaene, S., & Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. *Cognition*, *79*(1–2), 1–37.

Baars, B. J. (1997). *In the theater of consciousness: the workspace of the mind*. Oxford University Press.

Takeno: 10 features of consciousness

Husserl's 10 features of the functions of human consciousness

(1) **First-person property**: The sense that one is performing all things, i.e., a belief in the existence of the self or mind–body monism.

(2) **Orientation**: Orientation means that consciousness is always directed toward something.

(3) **Relationship between action and result**; **duality** of self- consciousness: Duality of self-consciousness means being aware of oneself.

(4) **Expectation**: Humans are always predicting the immediate future.

(5) **Function of determination and conviction**: Belief in the existence of things.

(6) **Embodiment**: Embodiment is the feature that the body is part of the self. All of us are conscious that our body is part of our self.

(7) **Consciousness of others**: The feature that enables us to discriminate our self from others.

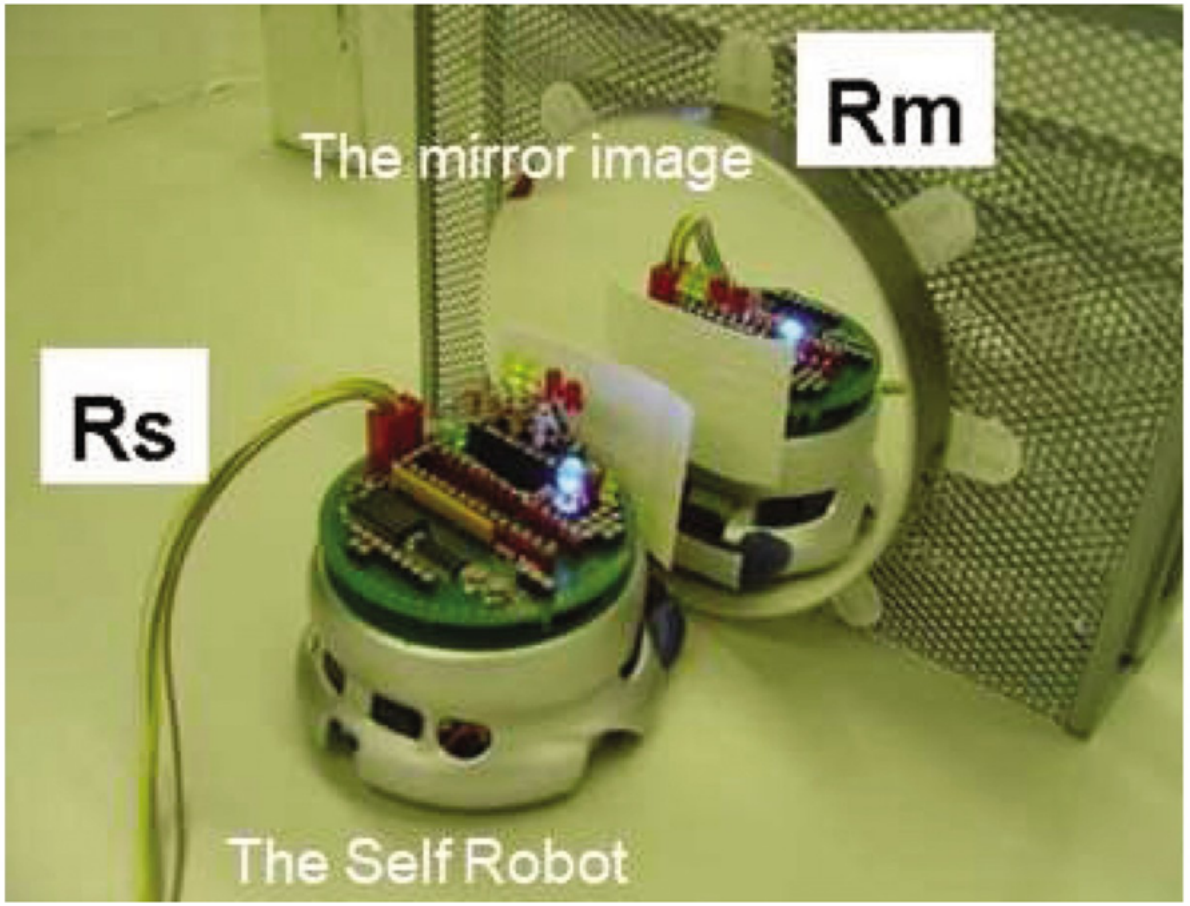
(8) **Emotional thought**: Reason is related to emotion and feelings.

(9) **Chaos**: Consciousness is ceaselessly out of balance.

(10) **Emotion**: Qualia of consciousness. The human senses of taste, hearing, smell, color, pain, etc., are deeply related to qualia.

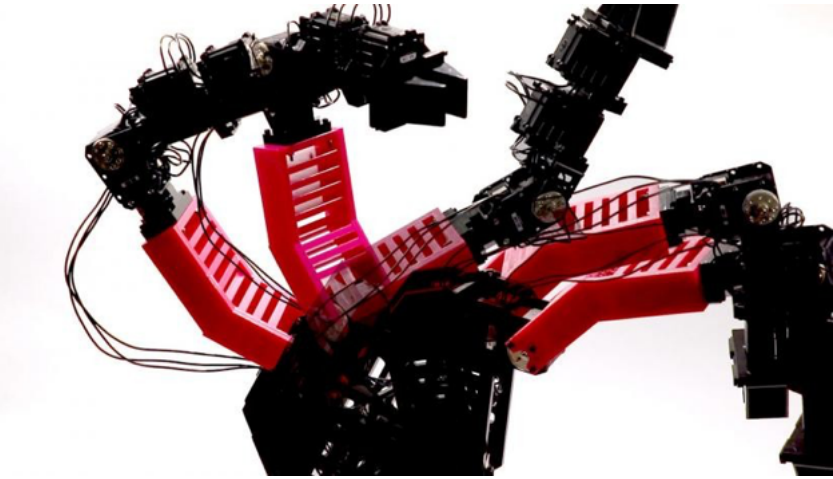
Takeno, J. Self-Aware Robots. 2nd Edition, Jenny Stanford Publishing (2022).

Takeno’s robot looking at itself in a mirror



Lipson’s robots know themselves

- Lipson at the Creative Machines Lab at Columbia University is creating a machine that will have "consciousness on par with a human".
- Lipson definition of consciousness:
  - “The capacity to imagine yourself in the future”
- His words:
  - “This will eclipse everything else we've done”



- These Self-Aware Robots Are Redefining Consciousness

  https://www.youtube.com/watch?v=chukkEeGrLM

Chen, B., Kwiatkowski, R. Vondrick, C. and Lipson, H. Fully body visual self-modeling of robot morphologies. Science Robotics (2022), 7, 68, pp. eabn1944.

Many other “conscious” robots

- Holland
- Chella
- Haikonen
- Hernández
- Kitamura
- Lanillos
- Tani
- …

What is a conscious robot?

- From the **outside**:
  - A robot that performs feats that are “associated to consciousness”.
  - For example: Mirror recognition or verbal report.
- From the **inside**:
  - A robot that “implements a theory of consciousness”.
  - For example: GNW or Orch OR.

- **A problem**: What is consciousness? A complex phenomenon difficult to detect in humans (neural correlates of consciousness, NCC).

The many forms of consciousness

- [Anthony] “Phenomenal consciousness, access consciousness, state consciousness, creature consciousness, introspective consciousness and self-consciousness”. Consciousness is a mongrel concept.

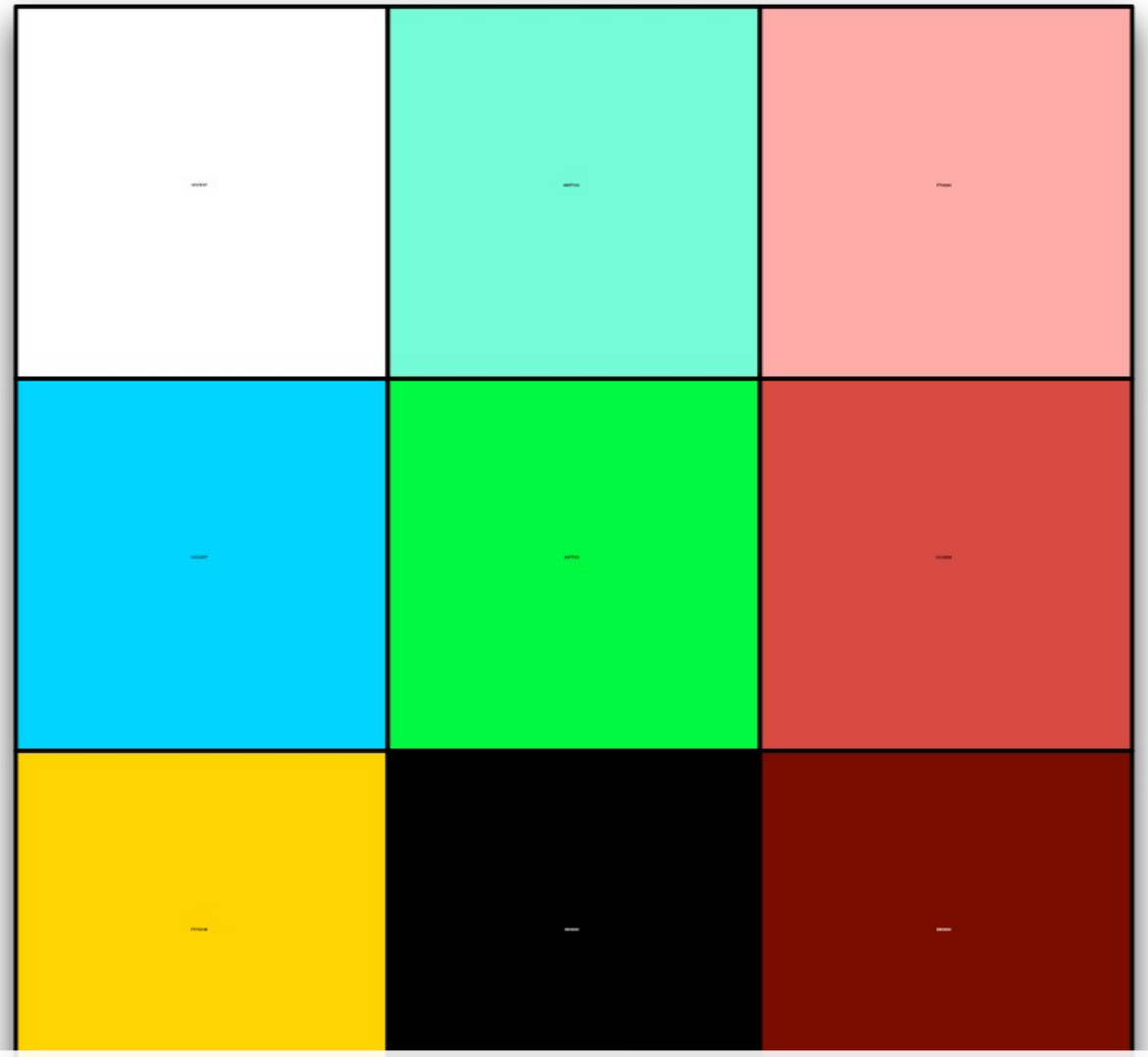
- From the perspective of robotics there are some major classes:
  - **Access** (Epistemic)
    - **External** (the **world** including other **agents**)
    - **Internal** (the **self**)
  - **Sentience** (Phenomenal)

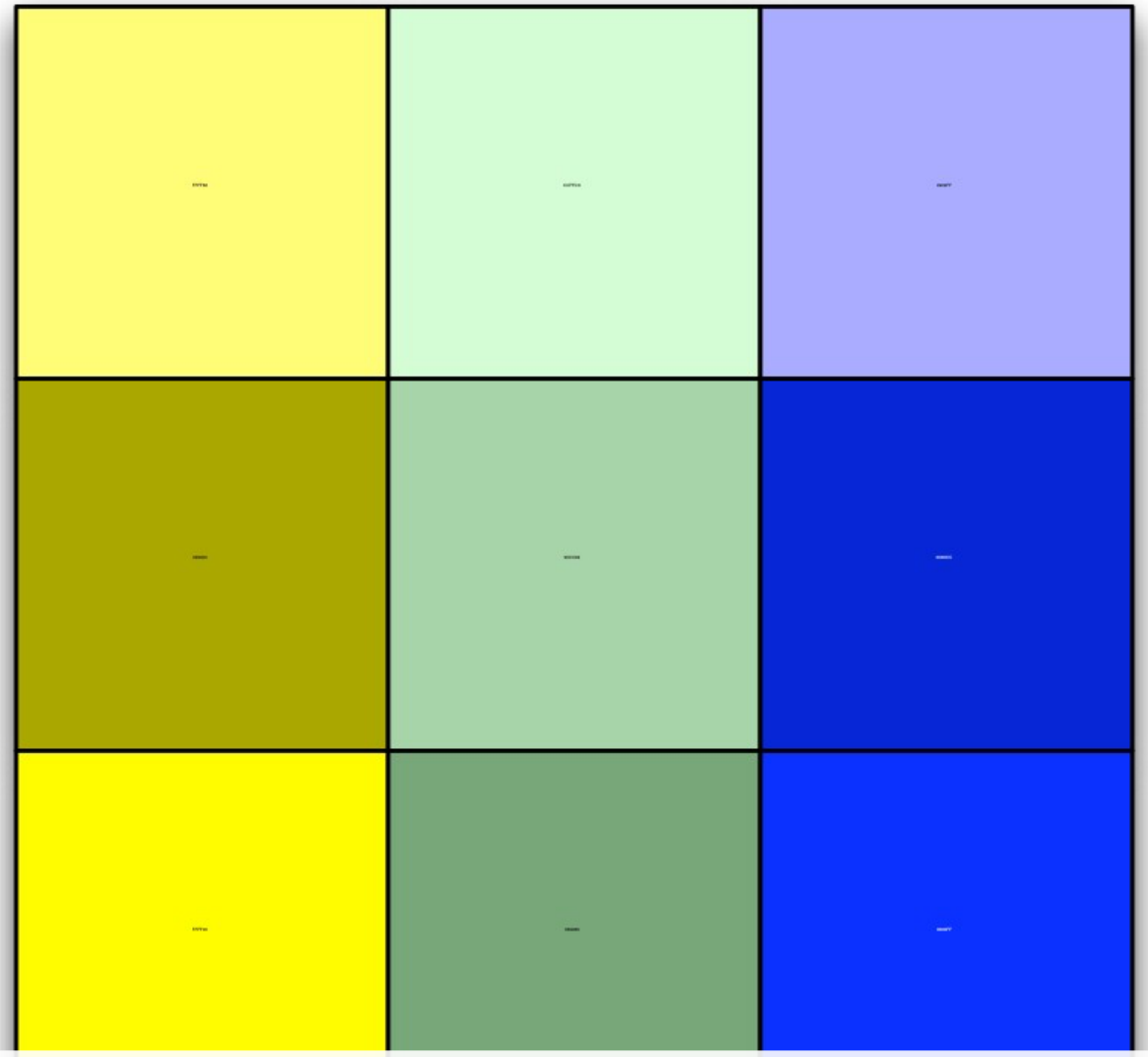
data & qualia

qualia

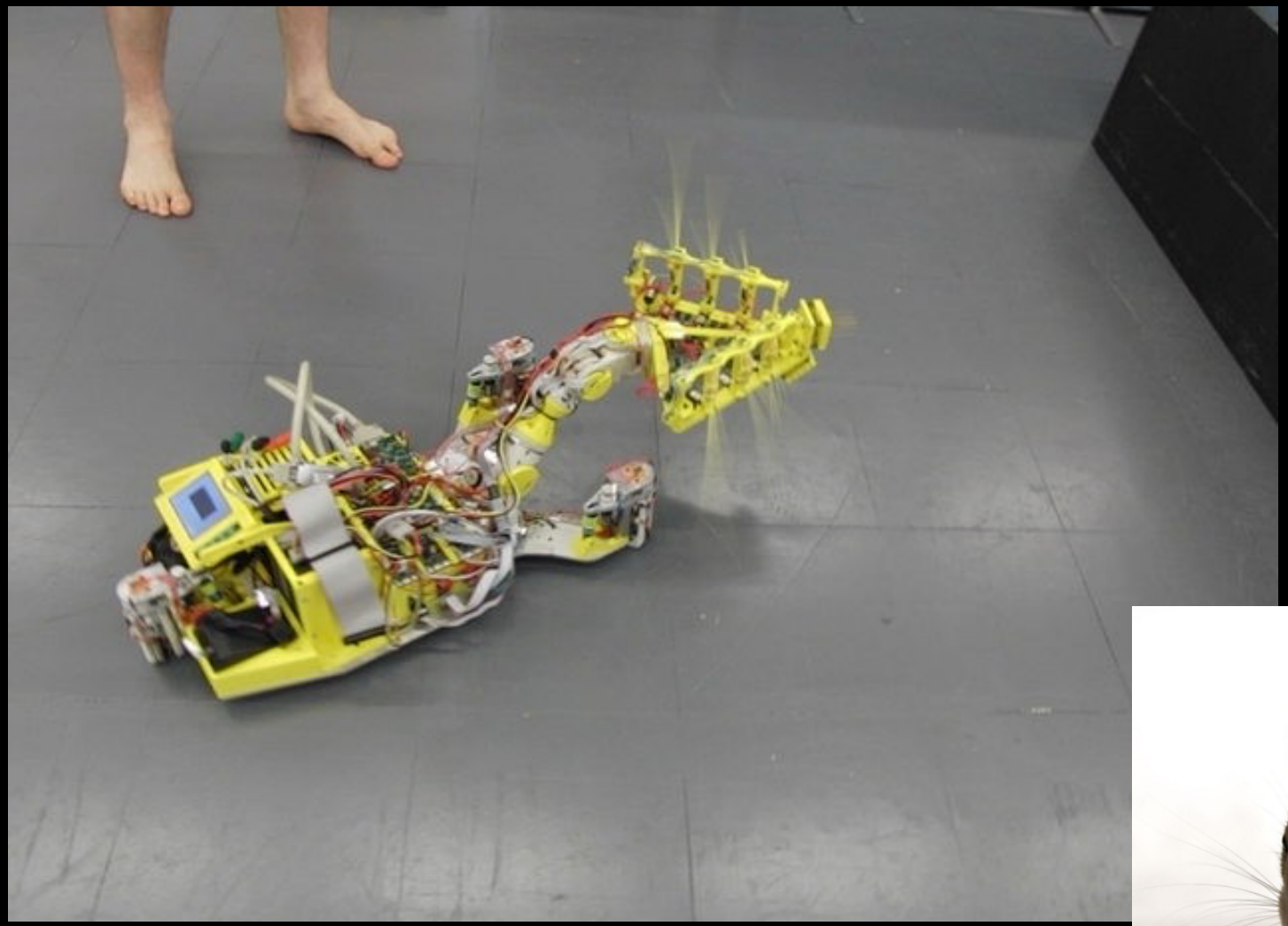
Sloman, A. (2007). Why some machines may need qualia and how they can have them: Including a demanding new Turing test for robot philosophers. *AAAI Fall Symposium*, *FS-07-01*, 9–16.

Photo by Mohamed Nohassi on Unsplash





The world of mice and men



FFFFFF	66FFCC	FF9999	FFFF66	CCFFCC	9999FF
00CCFF	00FF33	CC3333	999900	99CC99	0000CC
FFCC33	000000	660000	FFFF00	669966	0000FF

The world of roborats and philosophical zombies?

Theories of consciousness

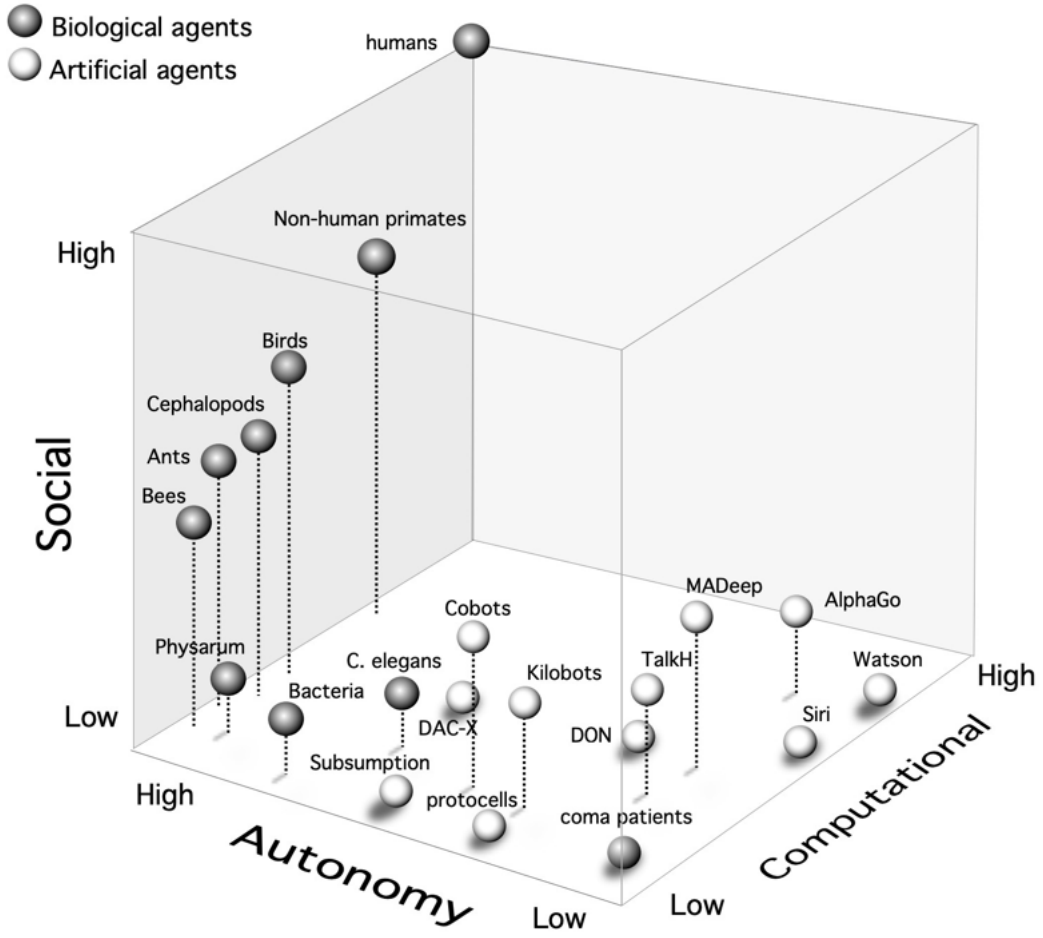
- So many theories of human consciousness.
- Some have major traction (neuroscience, neuro-robotics, quantum):
  - Global Workspace Theory (GWT).
  - Global Neuronal Workspace (GNW).
  - Higher-order thought (HOT).
  - Integrated Information Theory (IIT 4.0).
  - Recurrent Processing Theory (RPT).
  - Orchestrated objective reduction (Orch OR)
- In recent papers, up to 36 theories are identified:

Neuroscience
Psychology
Philosophy
Cybernetics
Religion
….

Signorelli, C. M., Szczotka, J., & Prentner, R. (2021). Explanatory profiles of models of consciousness - towards a systematic classification. *Neuroscience of Consciousness*, *2021*(2).
Seth, A. K., & Bayne, T. (2022). Theories of consciousness. *Nature Reviews Neuroscience*.

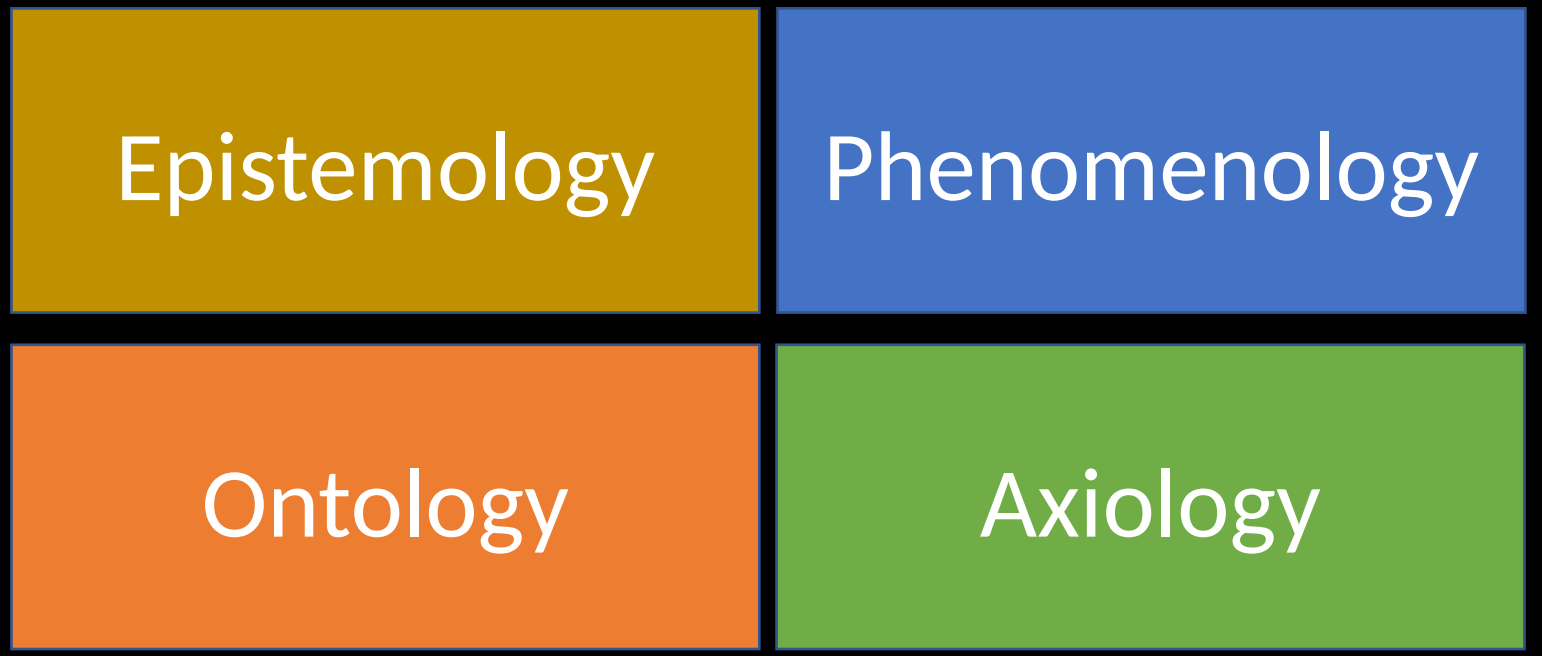
Complexity for consciousness

	Autonomic	Computational	Social
Building Blocks	Sensors, Actuators	Neurons, Transistors	Individual Agents
Systems-Level Realizations	Prokaryotes, Autonomic Nervous System, Bots	Cognitive Systems, Brains, Microprocessors	Population of Agents, Social Organizations
Emergent Phenomena	Self-Regulated Real-Time Behavior	Problem Solving Capabilities	Signaling Conventions, Language, Social Norms, Art, Science, Culture



Arsiwalla, X.D.; Solé, R.; Moulin-Frier, C.; Herreros, I.; Sánchez-Fibla, M.; Verschure, P. The Morphospace of Consciousness: Three Kinds of Complexity for Minds and Machines. NeuroSci 2023, 4, 79–102

Four philosophical viewpoints

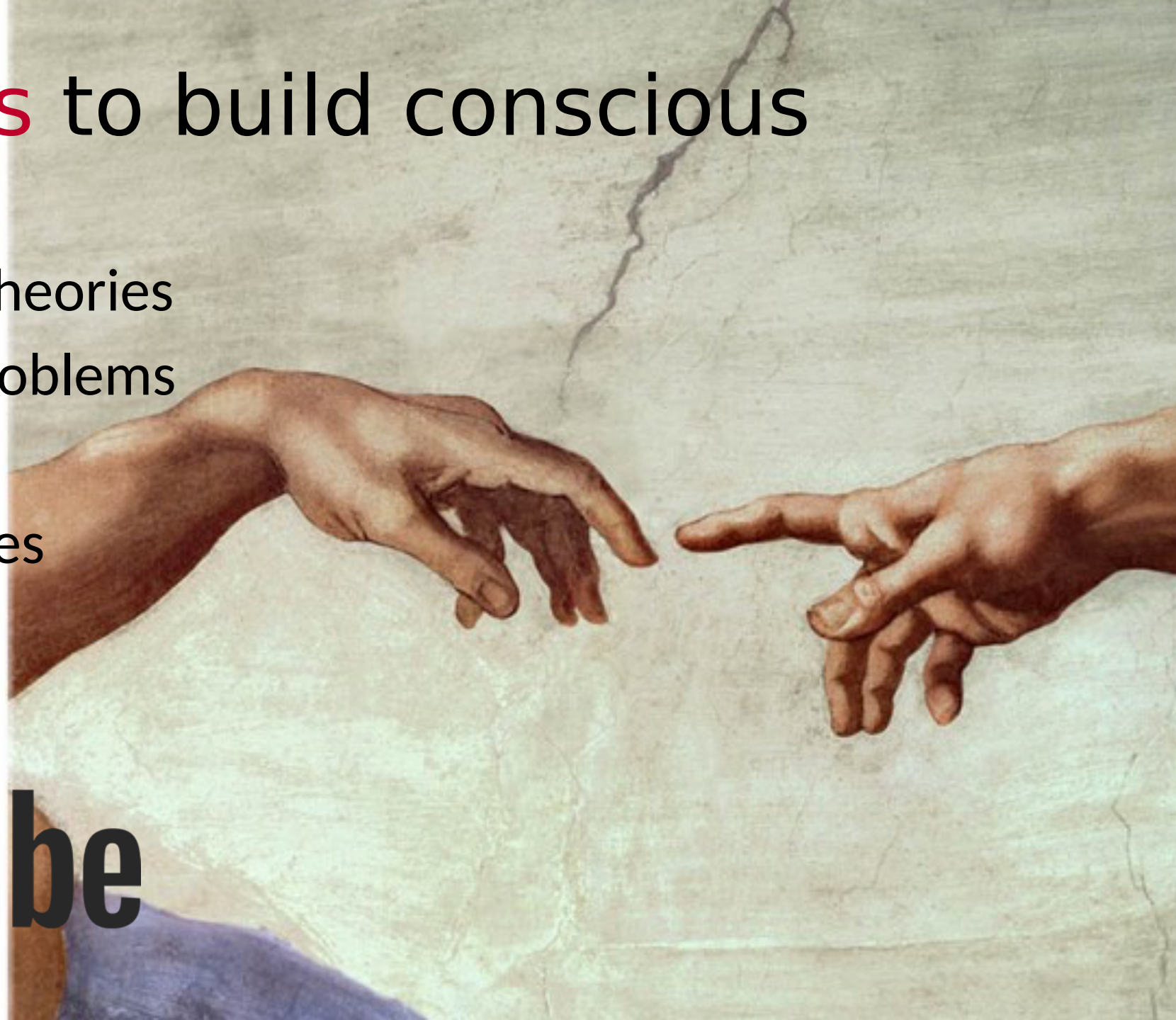


Assessment

Problems, Fallacies, Roadblocks

Four reasons to build conscious robots

- Testing biological theories
- Solving peoples' problems
- Playing God
- Doing show bussines

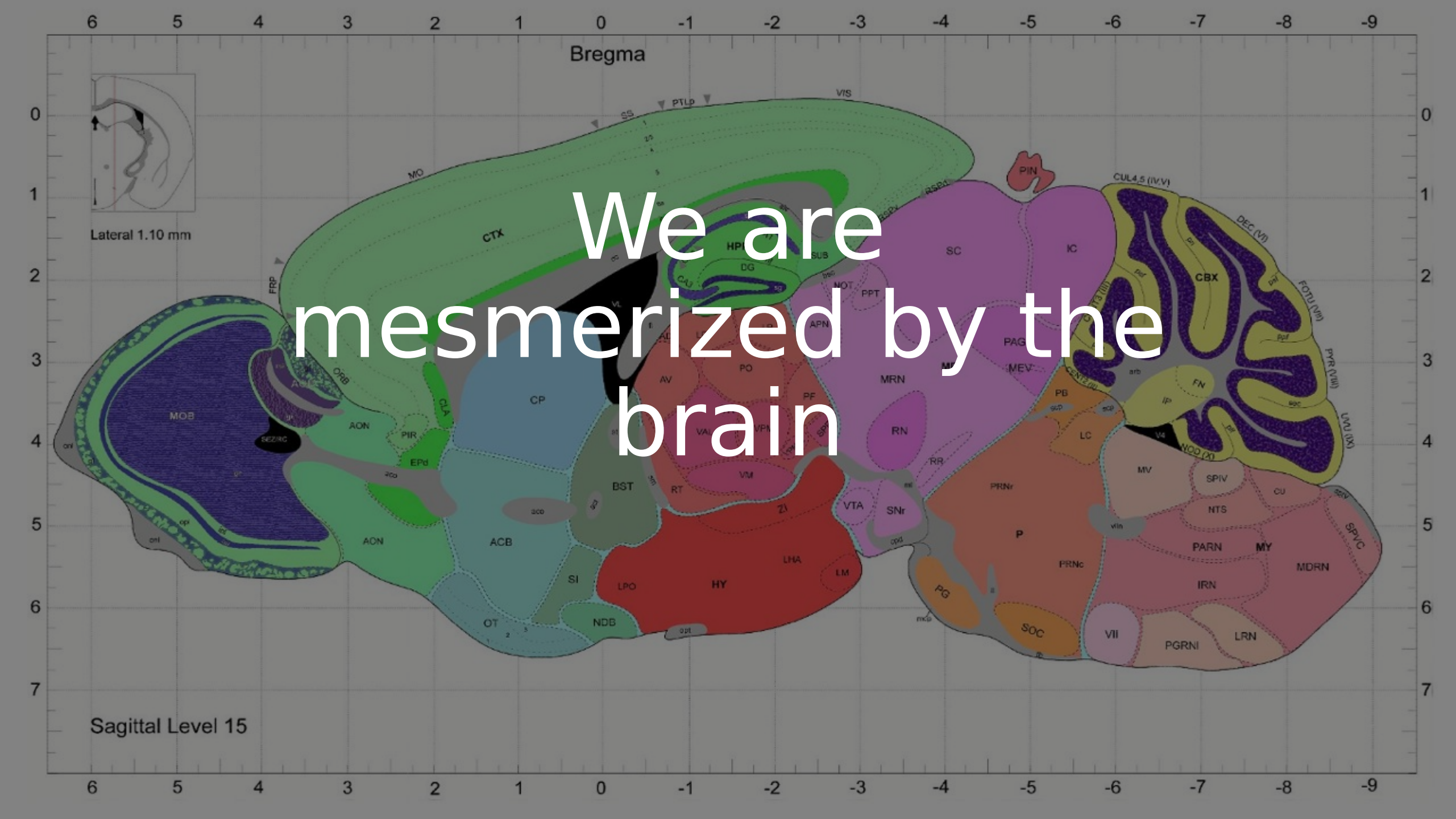


YouTube

Cats don’t recognise themselves in mirrors

Are cats conscious of themselves?

biologism can be misguiding


Bregma
Lateral 1.10 mm
Sagittal Level 15
We are mesmerized by the brain
CTX
MOB
AON
PIR
EPd
CP
ACB
BST
SI
OT
NDB
HY
LHA
LPO
LM
ZI
VTA
SNr
RT
VM
AV
PO
PF
MRN
RN
SC
IC
PAG
MEV
APN
PPT
NOT
HPF
DG
SUB
PIN
CBX
CUL4,5 (IV,V)
DEC (VI)
FOTU (VII)
PYR (VIII)
UVU (IX)
NOD (X)
FN
IP
MV
SPIV
NTS
PARN
MY
MDRN
IRN
SPVC
PRNr
PRNc
P
PB
LC
PG
SOC
VII
PGRNI
LRN
CU
6 5 4 3 2 1 0 -1 -2 -3 -4 -5 -6 -7 -8 -9

Fallacies in cognitive/conscious robotics

Cognition means thinking like a human

An AGI shall be human-like AI

The only viable architecture is a neural network architecture

Learning is necessary for suitable behaviour

Embodiment, situatedness, enaction

Uniqueness, sequentiality of consciousness

Some major (ongoing) debates

Consciousness vs cognition

Consciousness vs awareness

Hard problem of consciousness

Emergent AI consciousness

Biological vs machine consciousness

Machine consciousness ethics

“Machines cannot feel”

- Turing’s argumenta against the idea that machines can’t actually think because they will never be possible to build a machine that can do such-and-such:
- Mysterians and other families (biochauvinists, negationists, etc)
- Complexity/Emergence/Autopoiesis
  - Absent Qualia (AQ) arguments
  - Zombies
    - Chinese Nation
  - …
- As we have seen in the history of AI this is a receding horizon problem.

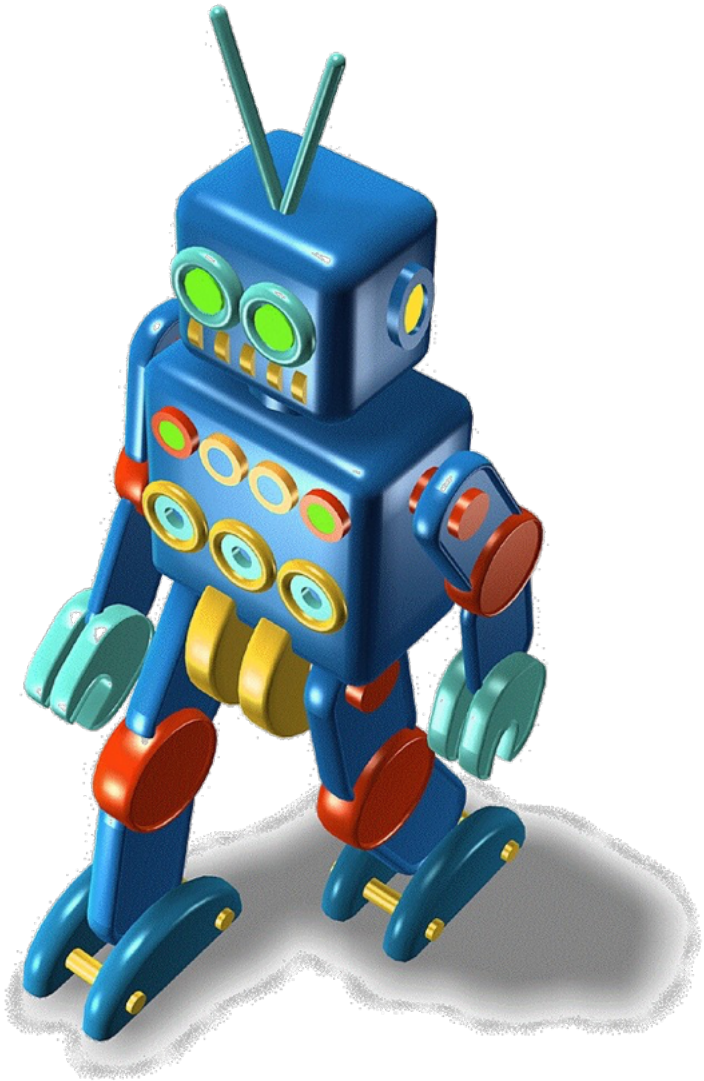
Robot consciousness and ethics

- *“Consciousness is what matters on a moral scale”*
- A red herring for robot builders
- Metzinger: “We shall forbid building systems that can suffer”.
- Doomsayers: “AI is going to kill us all when aware”.

- Science and technology development is **ethics-neutral**.
- Science and technology is risky when used. Use with care.
- Please stop bothering.

Building Consious Machines

Two Horizon Europe Projects



Two active projects

Funded by the European Union

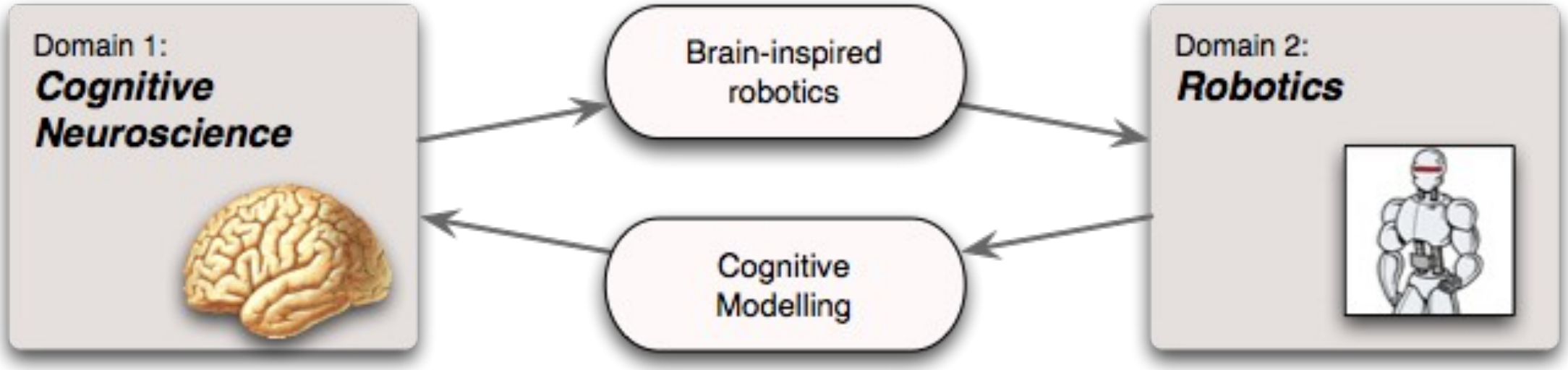
The construction of conscious robots enables two things:

- The exploration of theories of self-awareness and associated ***bodily*** phenomena:
  - This is what we do in **METATOOL**, exploring the self-awareness aspects in the metacognitive control of tool use and invention.
- The improvement of performance and resilience thanks to better ***understanding*** of the world and itself:
  - This is what we do in **CORESENSE**, exploring general theories of consciouness beyond the biological realm to create adaptive, resilient robots capable of self control.

MetaTool
Robots inventing tools

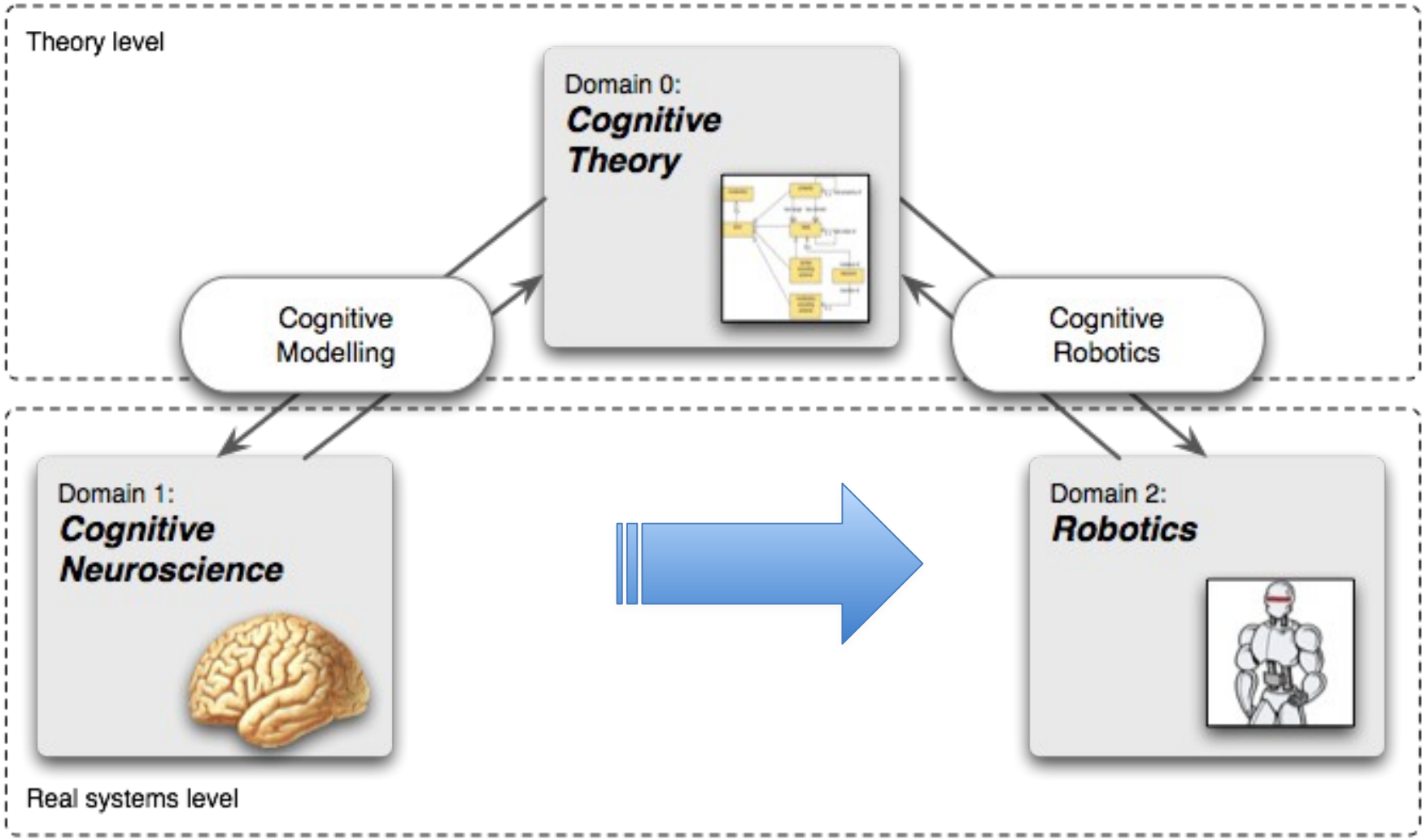
CORESENSE

Beware the constructive mismatch



attention?

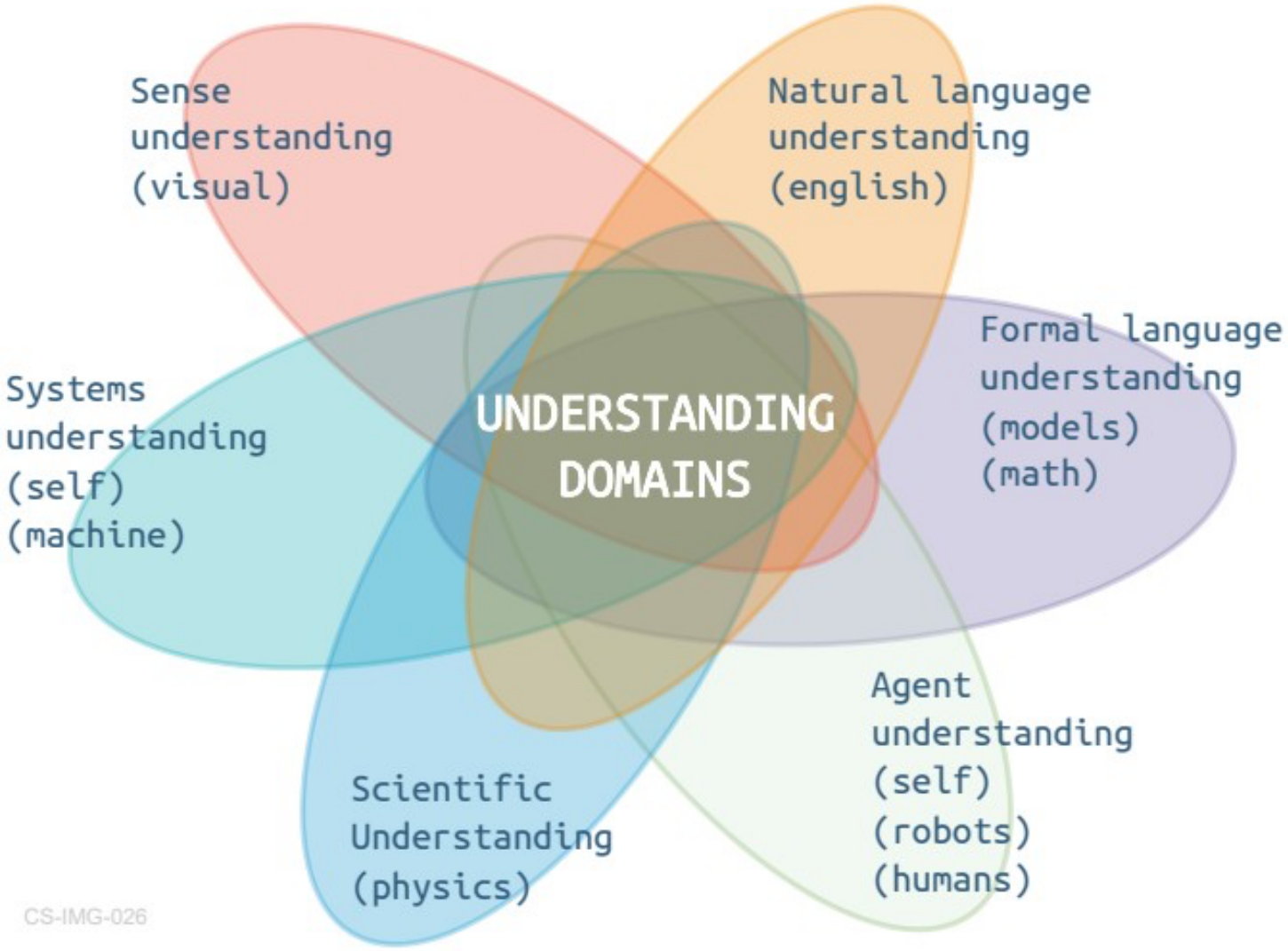
Build theories first



Project purpose

Develop a **cognitive architecture** for **deep understanding**

Understand what?



Three testbeds

CORESENSE

social

alignment

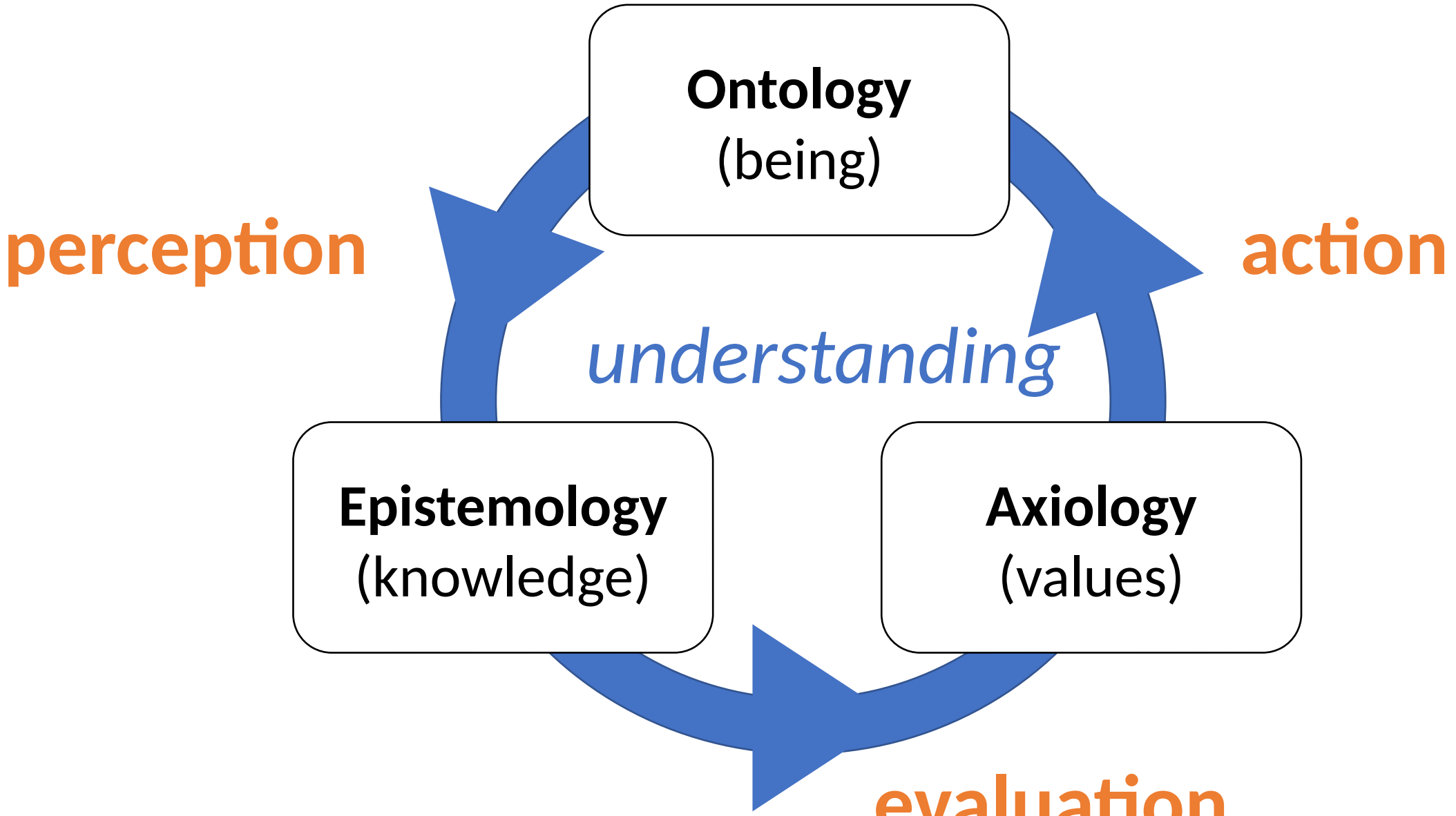
aerial

resilience

factory

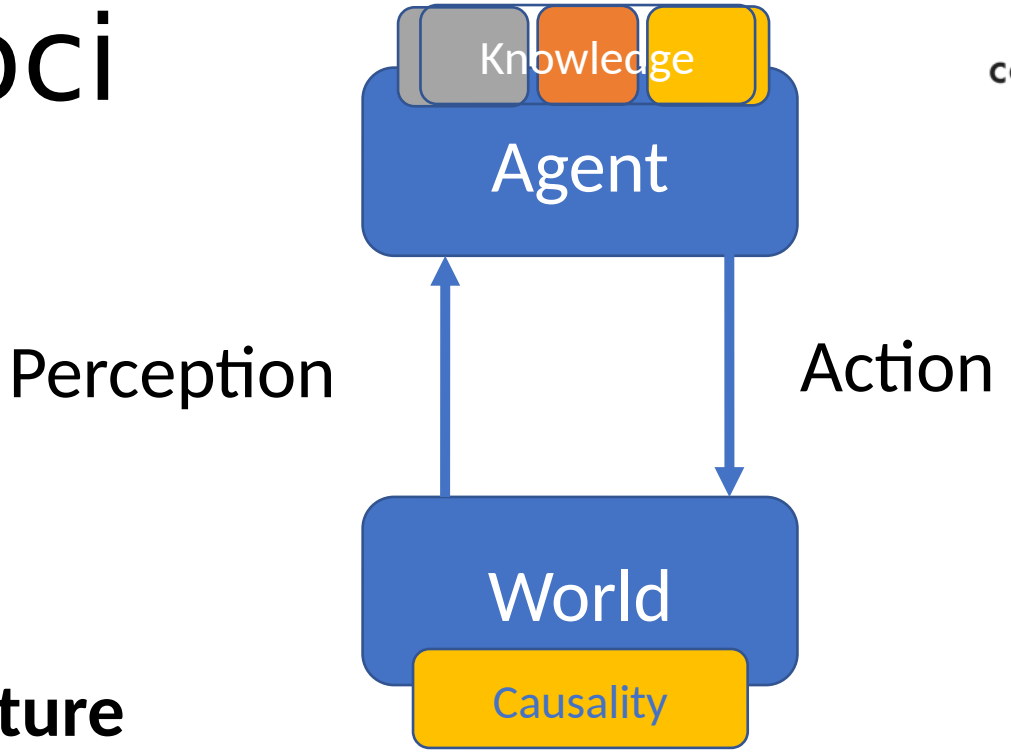
flexibility

CORESENSE


Ontology
(being)
perception
action
understanding
Epistemology
(knowledge)
Axiology
(values)
evaluation

Two fundamental foci

- **Understanding**
  - Making sense of what is “perceived”
- **Awareness**
  - Continuous sense-making of the world

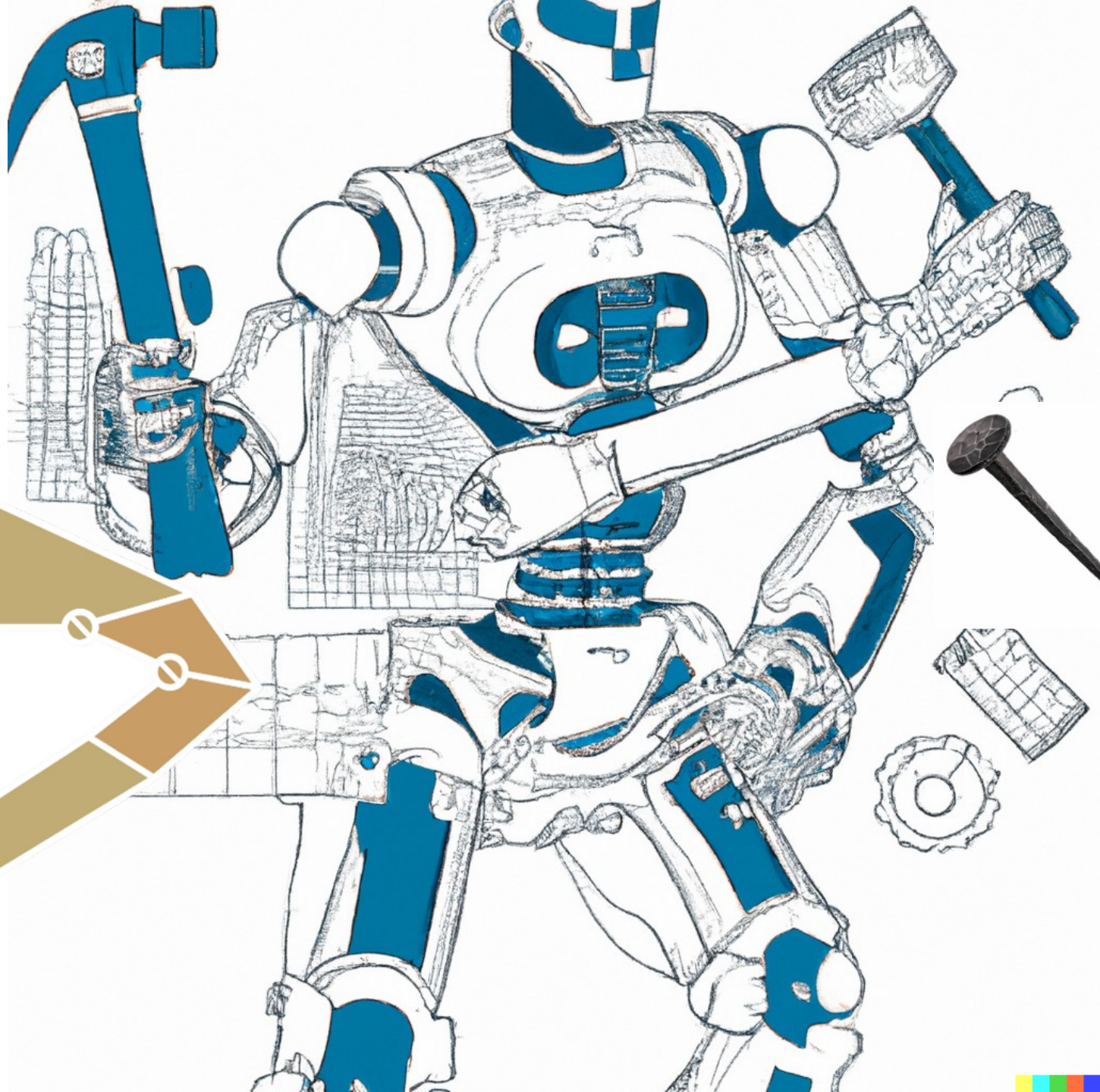


- CoreSense is a “**cognitive” architecture**
  ✉ based on **knowledge**

Agent knowledge is in **sync with the world at a deep level**

Knowledge is leveraged in producing **meaningful action**

Metacognition in tool invention



MetaTool

Robots inventing tools

When the **body** is **extended by a tool**
it becomes part of it

The body image changes

The tool moves from the scope of **awareness** into the scope of **self-awareness**

The robot shall be “aware” of the whole system:
Body-Tool-Environment

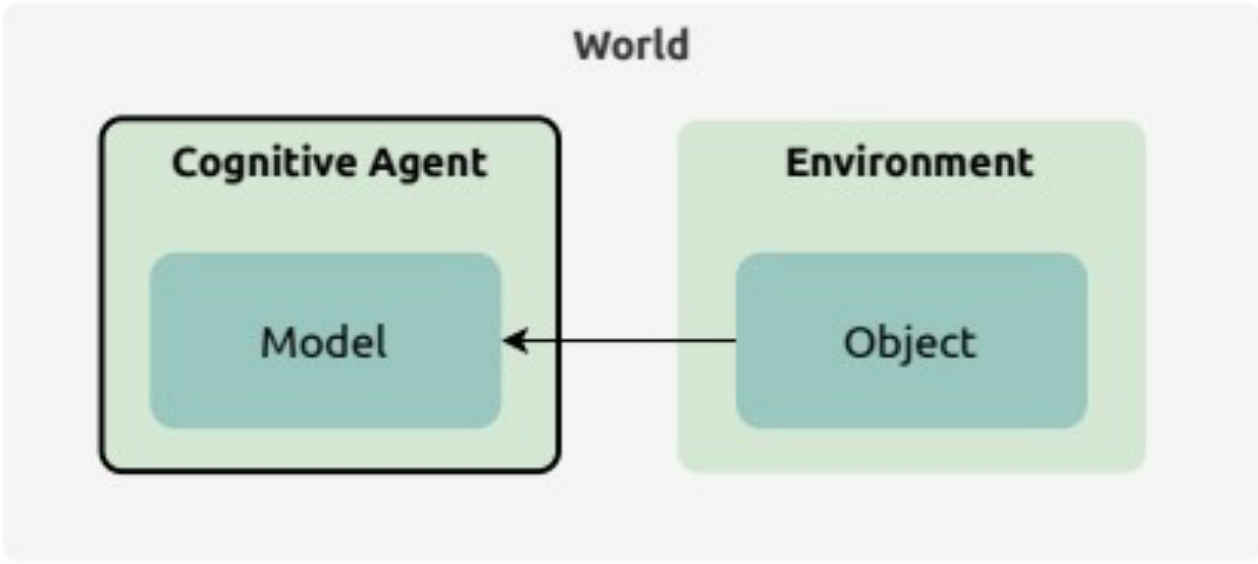
We investigate **how robots can extend themselves with tools** and the role that **awareness** – world, tool and self– and **metacognition** do play in this.

We take inspiration from very **ancient tool use and tool making**

Towards a Concept of Awareness

- There are ontologic, epistemologic, phenomenologic, ontologic and axiologic aspects of awareness:

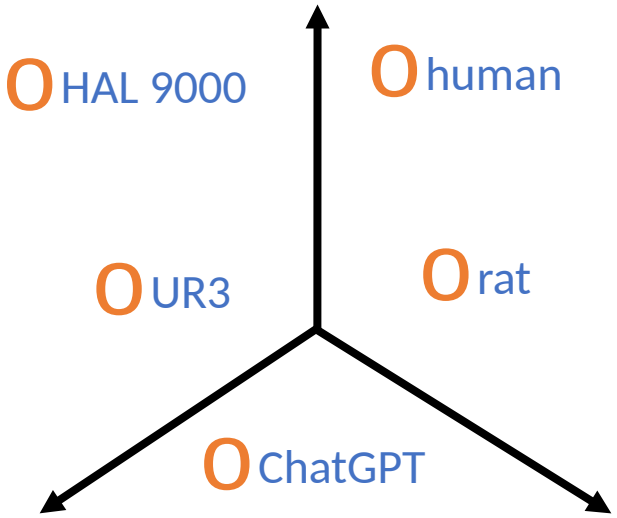
  - What is there
  - What is known
  - What is important
  - What is felt



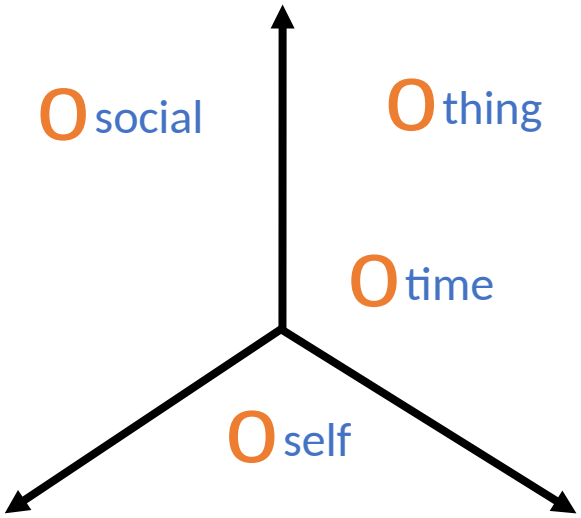
CS-IMG-034

The many Dimensions of Awareness

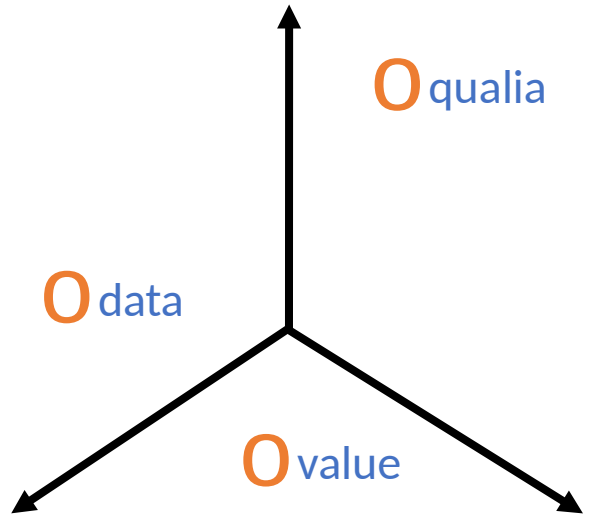
Subject of Awareness



Object of Awareness



Nature of Awareness

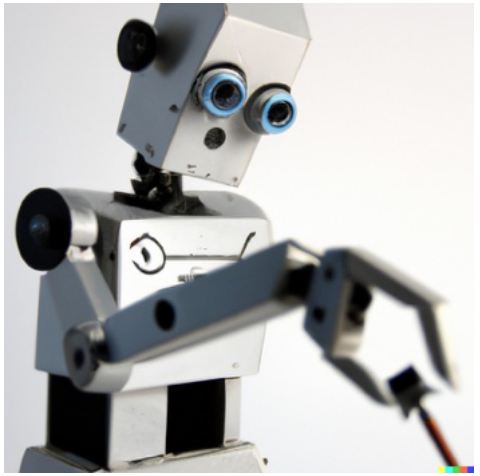


awareness: real-time understanding of sensory flows

Have actionable, mission critical mental models, representations, beliefs that something is so.

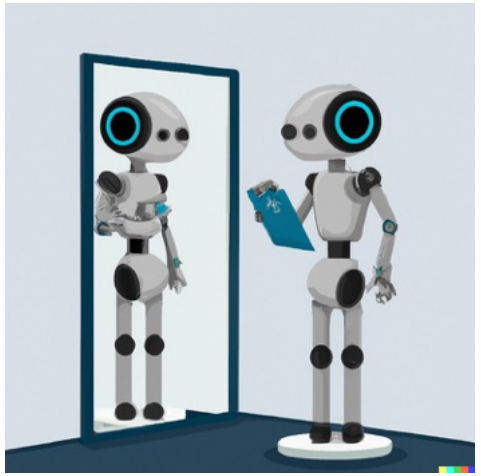
Beliefs (models), desires(missions), intentions(exertions) + real-time

use tools



invent tools

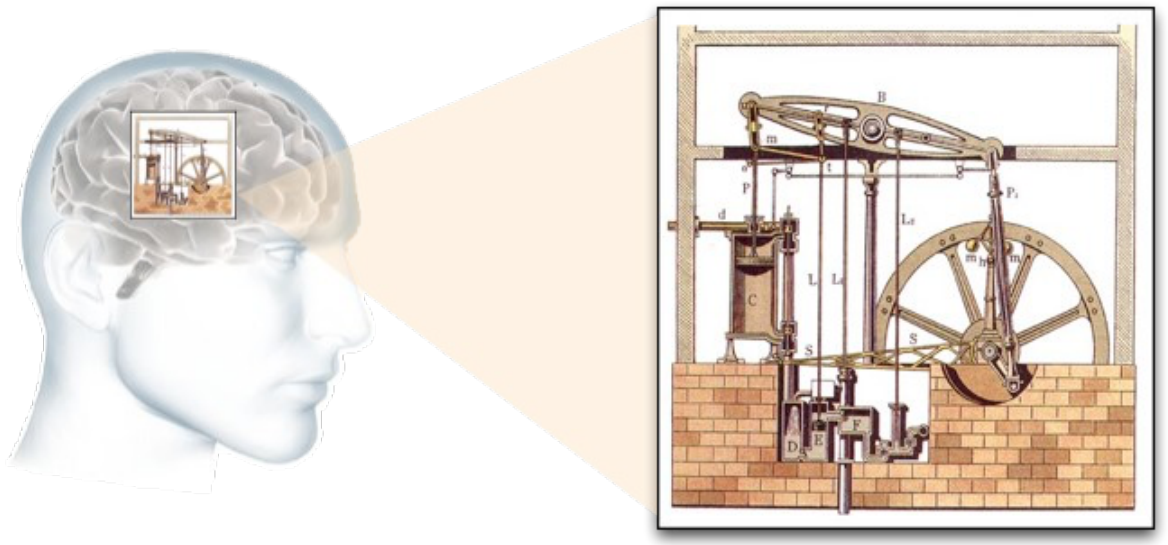
self-awareness



abstraction

world domination

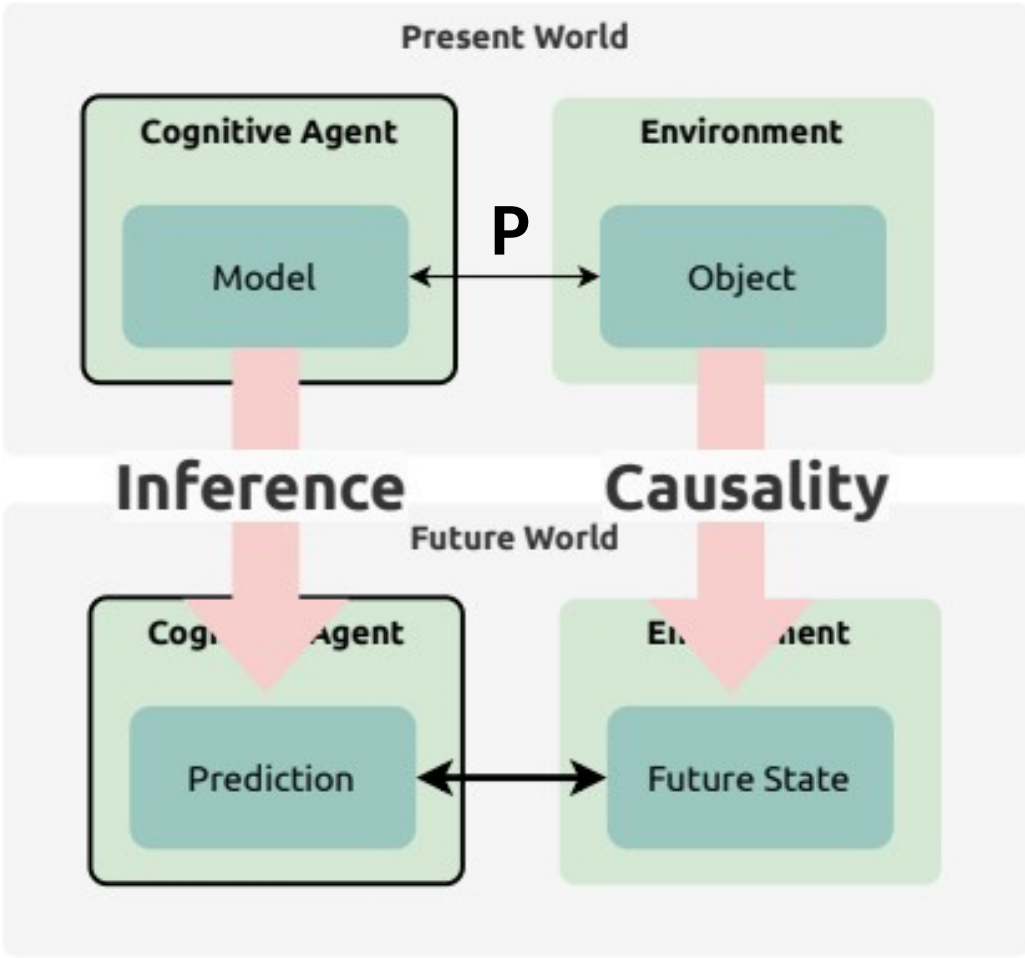
The agent models the world



Models are not just photographs.
They shall be **functionally equivalent** dynamical systems.

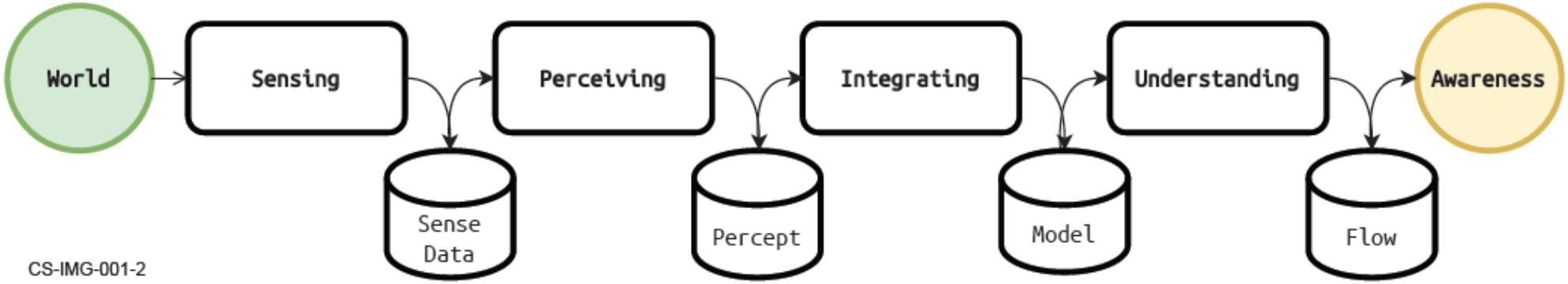
Definition of understanding

“A subject **S**
understands a phenomenon **P**
if it has a set of models **M** of **P**
and those models can be executed
to make valid inferences about the
phenomenon”



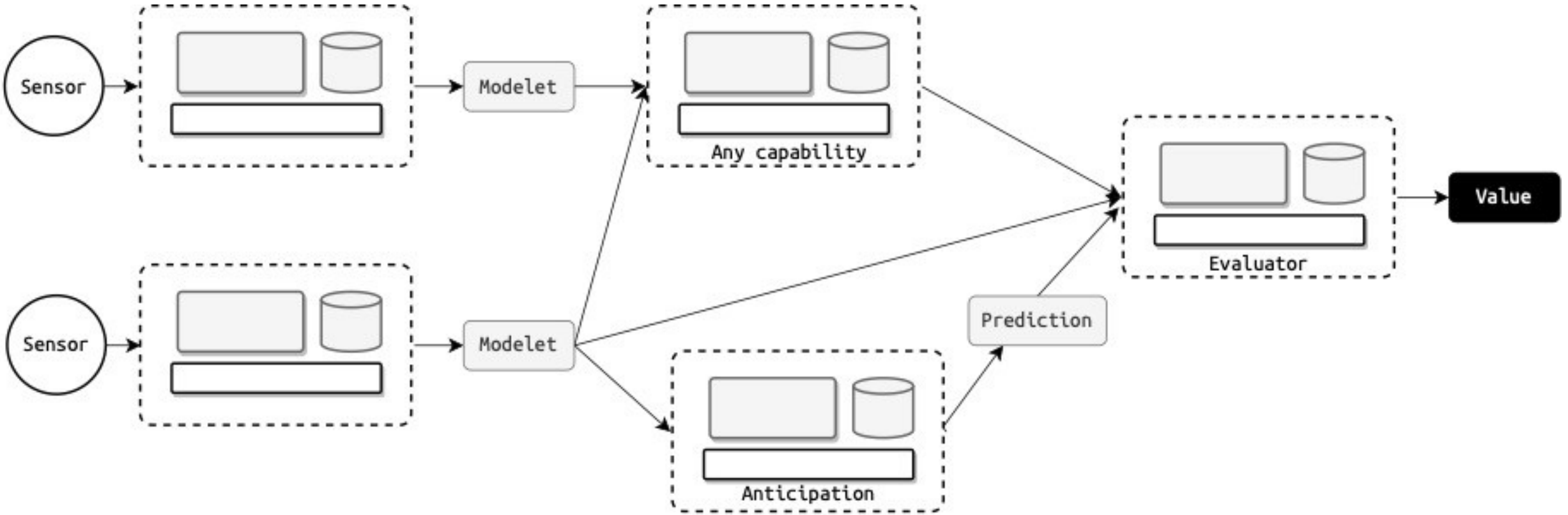
CS-IMG-035

Awareness



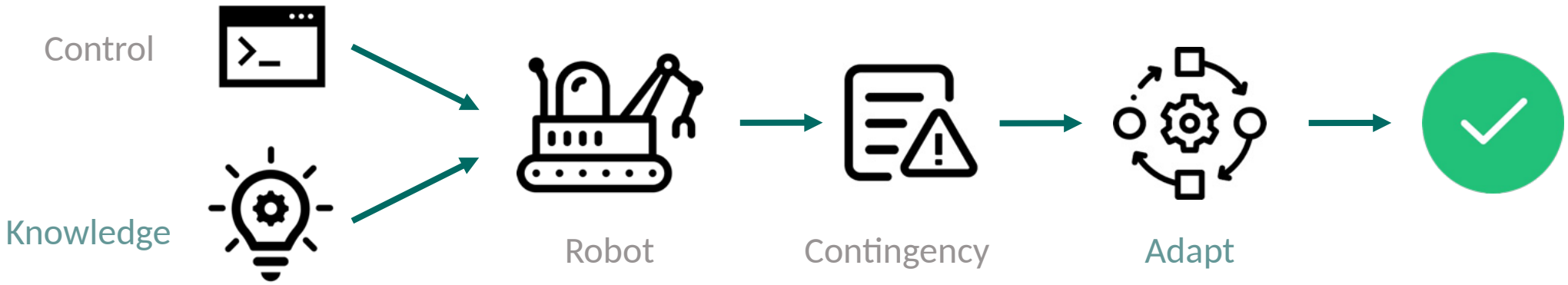
- **meaning:** understanding for a particular agent with regard to the agent's goals in a specific context.
- An agent (subject) is **aware** when it is continuously computing meaning from phenomena (object)
- **awareness**: real-time understanding of sensory flows for a particular agent with regard to the agent's goals in a specific context.

Awareness and Value



A Concrete Example

- Endow robots with a **better understanding about what is happening**, **what capabilities it has**, and what to do to **reach its goals**.
- **Exploit deep system-architectural knowledge** to reason at runtime about crucial aspects that are explicit at design phase.



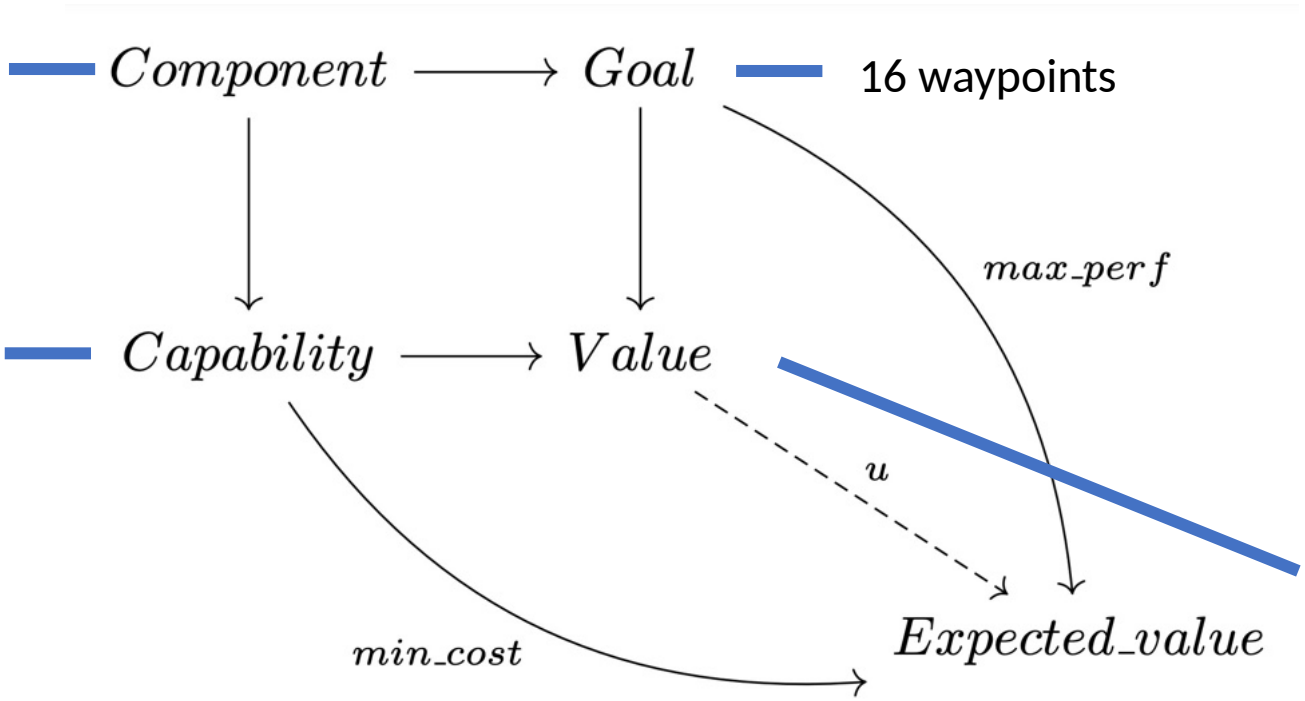
Esther Aguado

Adaptation in the The Marathon 2

Navigation2 ROS stack paradigm experiment. [Macenski et al., 2020]

AMCL, A* planner,
LiDAR, depth camera

Plan, Localize,
Control, Navigate



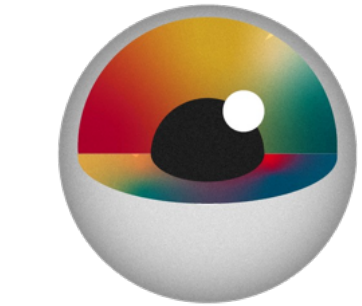
Reach waypoints
Time efficiency
Point accuracy
Energy consumption
Safety

Esther Aguado

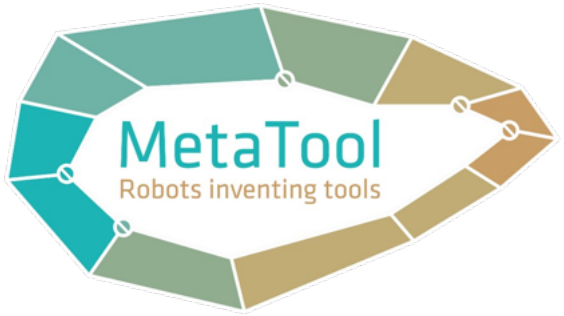
Want to track us?

coresense.eu

metatool-project.eu



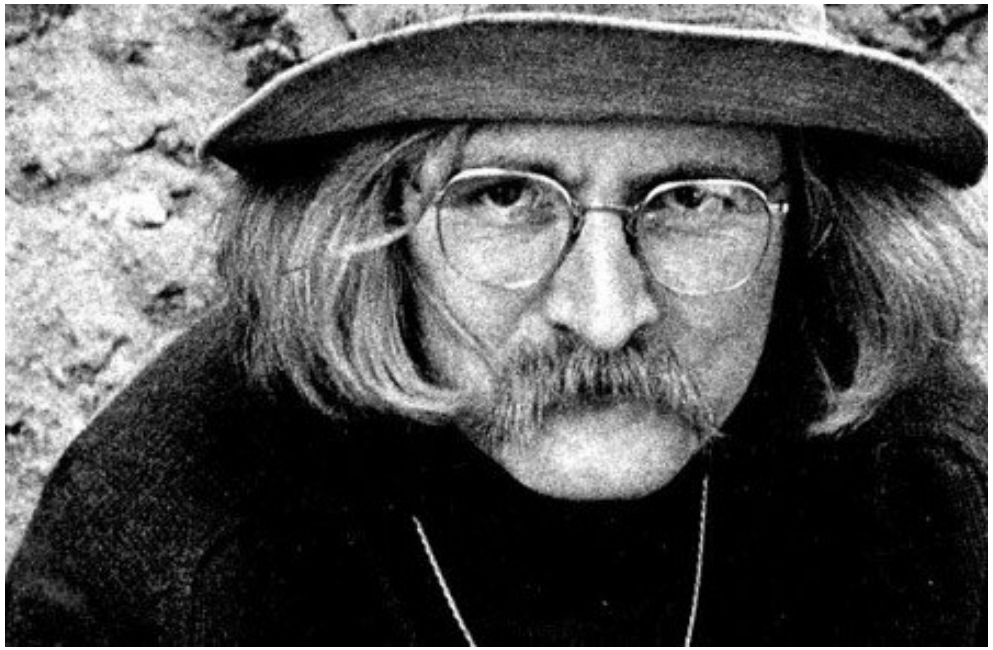
CORESENSE



Awareness Inside – EIC Pathfinder Challenge

“This conversation can serve no purpose anymore”

Richard Brautigan - 1967
All Watched Over by Machines of Loving Grace



I like to think
(it has to be!)
of a cybernetic ecology
where we are free of our labors
and joined back to nature,
returned to our mammal
brothers and sisters,
and all watched over
by machines of loving grace.

European Innovation Council

The METATOOL project has received funding from the European Innovation Council through the Pathfinder Challenges grant No. 101070940.

The CoreSense project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No. 10107054

Funded by the European Union