# Towards a Theory of Awareness

**Ricardo Sanz**

Universidad Politécnica de Madrid

**AI with Awareness Inside - AWAI 2024**
24 February, 2024 - Rome, Italy
Within the 16th International Conference on Agents and Artificial Intelligence - ICAART 2024

CORESENSE

Funded by the European Union

The CORESENSE project is funded by the EC Horizon Europe programme through grant No. 101070254.
The METATOOL project is funded by the EIC Pathfinder programme through grant No. 101070940.

MetaTool
Robots inventing tools

European Innovation Council

# AWAI Schedule

Schedule - ICAART 2024

**9:15 Large Language Models and Artificial Consciousness**

**10:45 AWAI Session 1**

Towards Value Awareness in the Medical Field. Manel Rodriguez-Soto

Towards a Definition of Awareness for Embodied AI. Giulio Antonio Abbo

Fine-Grained Clustering of Social Media: How Moral Triggers Drive Preferences and Consensus. Emanuele Brugnoli

An Ontology for Value Awareness Engineering. Andrés Holgado-Sánchez

AwarePrompt: Using Diffusion Models to Create Methods for Measuring Value-Aware AI Architectures. Kinga Ciupinska

*12:45 Lunch*

**14:15 AWAI Session 2**

Notes on Measures for Information Access in Neuroscience and AI Systems. Giulio Prevedello

Towards a Theory of Awareness. Ricardo Sanz

**15:00 Panel for the AWAI Special Session: AI with Awareness Inside**

# Content of paper

- Rationale for awareness in autonomous robots
- Rationale for a Theory of Awareness (ToA)
- A short survey of perspectives/theories of awareness/consciousness
- Characteristics of a scientific and operationalisable ToA
- Essential elements for a ToA
- First steps to a ToA
  - Domain
  - Concepts

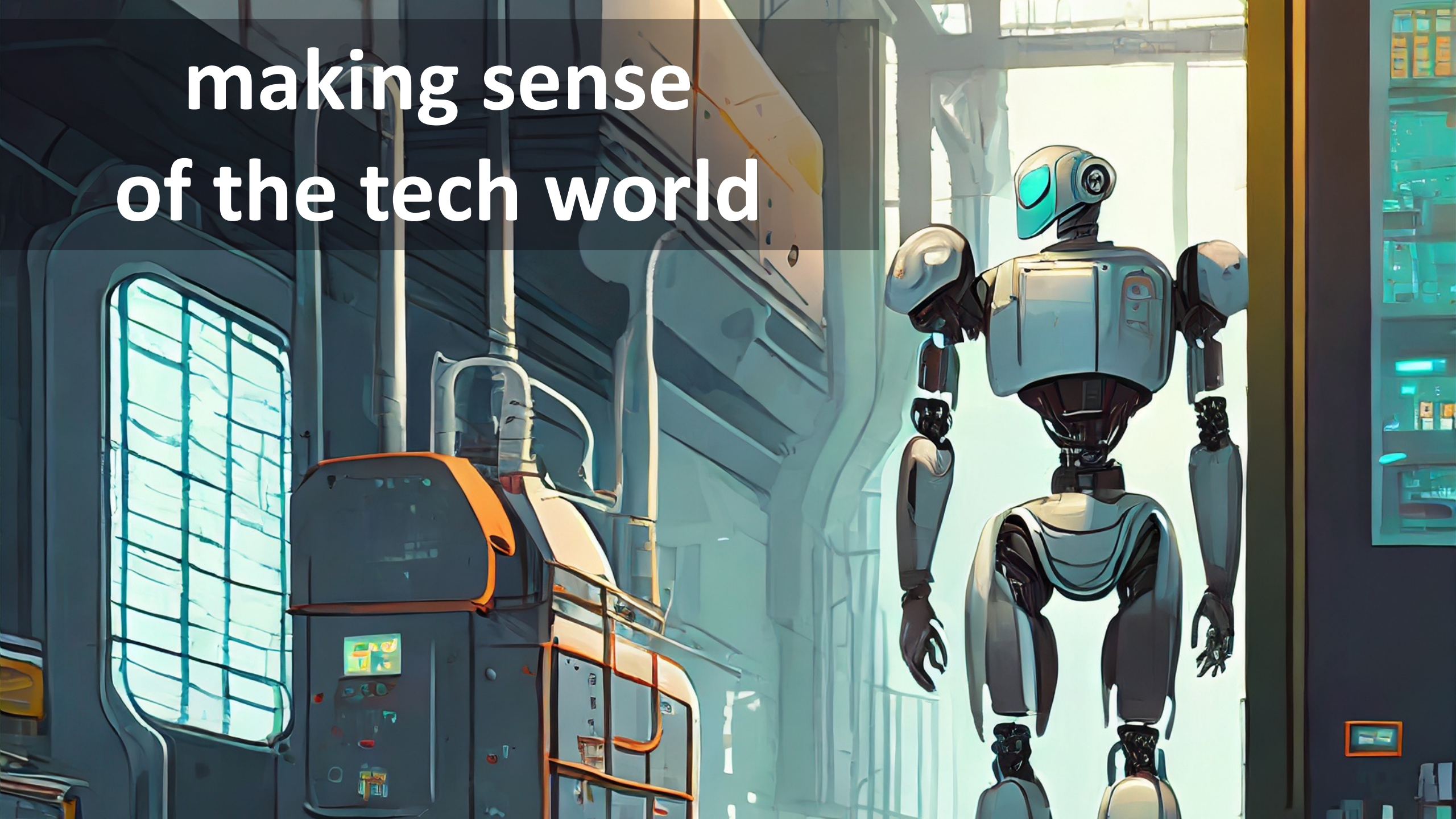
# Paper abstract

*Observation of humans and animals shows that* ***awareness is a critical aspect of mental processes for those agents that operate in changing environments****. Responding to potentially dangerous situations and leveraging environmental affordances are essential capabilities for autonomous agents' ecological viability. Agents need to be aware of their situations. Artificial autonomous systems construction depend on using suitable system architectures and applying proven engineering methods. While current systems display a certain degree of awareness, it is unclear what principles shall be used in their design. We are in a pre-scientific, pre-technological situation concerning awareness. Unfortunately, the scientific analysis of the awareness phenomena is quite difficult because its principles cannot be easily isolated in fully functioning human minds. We need a clean, formal theory of general awareness of universal nature. This theory should be applicable both to humans and machines, and not exclusively bound to the psychology and neurobiology of living animals. In this position paper, the authors argue for developing such a theory, state some requirements for it and propose an initial conceptual seed for a future theory of awareness that orbit around the idea that awareness is the real-time understanding of sensory flows.*

# Short Rationale

For a theory of awarneess


making sense
of the world


making sense
of the tech world

*We pursue an “awarenesss” technology to make better robots (or AIs)*

*Robots that see and understand.*
*Robots that can perform meaningful actions.*

*We ambition a clean, simple, formal, universal, operational,* ***theory of awareness***

*This theory should be applicable both to animals and machines.*

# Anthropomorphism in conscious LLMs

"Interactions with large language models (LLMs) have led to the suggestion that these models may soon be conscious. From the perspective of neuroscience, this position is difficult to defend. For one, the **inputs to LLMs lack the embodied, embedded information content** characteristic of our sensory contact with the world around us. Secondly, the **architectures of present-day artificial intelligence algorithms are missing key features of the thalamocortical system** that have been linked to conscious awareness in mammals. Finally, the **evolutionary and developmental trajectories that led to the emergence of living conscious organisms arguably have no parallels in artificial systems** as envisioned today. The existence of living organisms depends on their actions and their survival is intricately linked to multi-level cellular, inter-cellular, and organismal processes culminating in agency and consciousness."

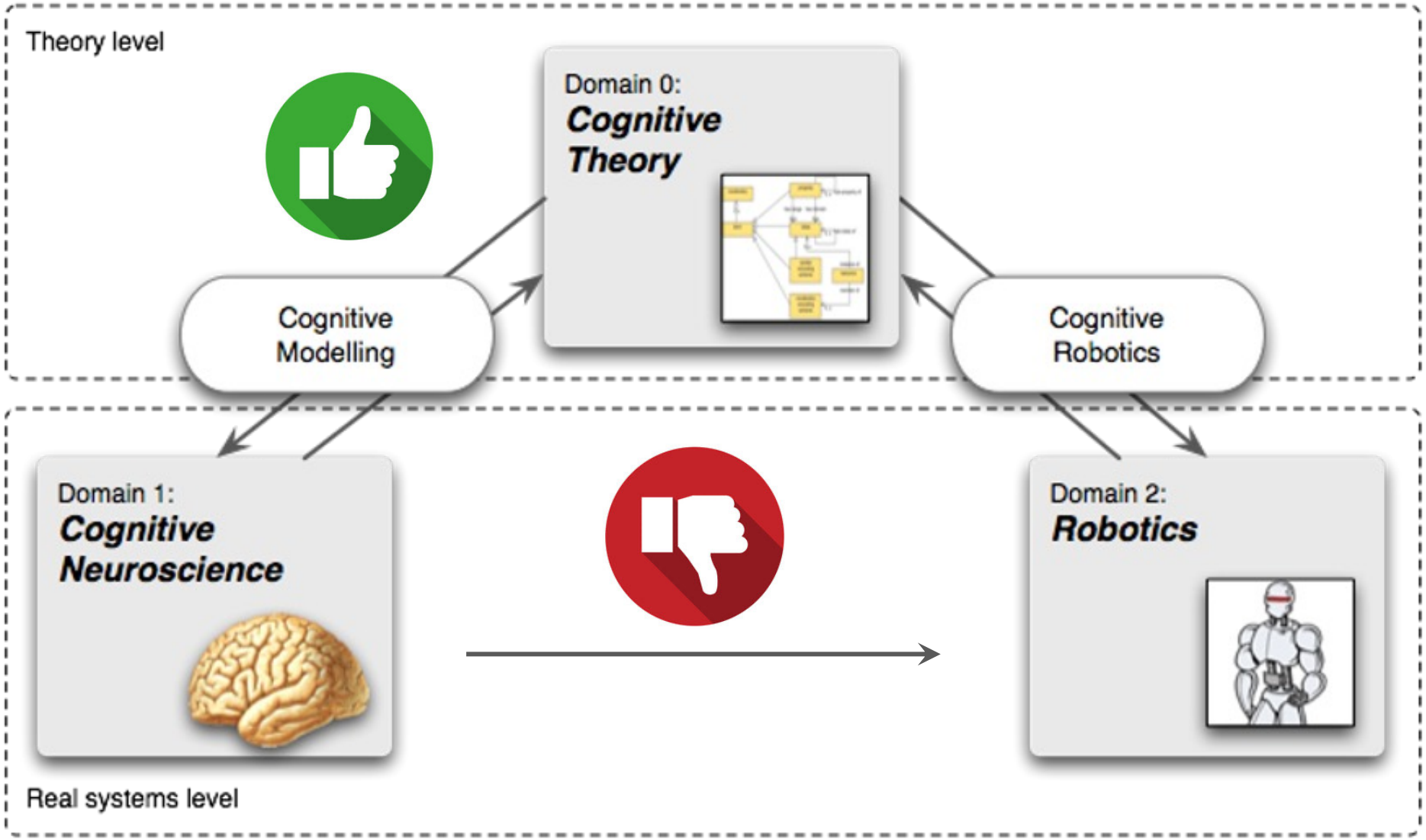
Sensors

Structure

Evolution

J. Aru, M. E. Larkum, and J. M. Shine, "The feasibility of artificial consciousness through the lens of neuroscience," *Trends Neurosci.*, pp. 1–10, 2023, doi: 10.1016/j.tins.2023.09.009.

# Two ways towards aware robots



# A theory of Awareness?

# Characteristics of a Scientific Theory

Empirical Basis

Scope

Consistency

Explanatory Power

Testability

Predictive Power

Consensus

Understandability

Simplicity

Reproducibility

Peer Review

# Elements for a Theory of Awareness

| | |
|---|---|
| **Domain** | A theory explains a specific natural phenomenon or a particular domain of inquiry. It defines the scope of what it seeks to explain. In our case: “awareness”. |
| **Concepts** | The theory contains a set of well-defined concepts that provide the vocabulary and framework for discussing and understanding the phenomenon. |
| **Hypotheses** | The theory often generates specific hypotheses, which are testable predictions or statements about how certain variables or factors are related within the defined domain, and guide empirical research. |
| **Laws or Principles** | A theory may incorporate scientific laws or fundamental principles typically derived from empirical data and observations, that describe relationships or patterns observed within the phenomenon. |
| **Causality** | The theory specifies causal relationships between the concepts and variables involved. |
| **Explanatory Power** | A theory should have a high degree of explanatory power, meaning it can account for a wide range of observations and data within its domain. |
| **Predictive Power** | A strong theory can make accurate predictions about future observations or experiments, i.e. should be verifiable through empirical testing. |
| **Models** | In some scientific theories, especially in the physical sciences and engineering, formal models may be used to describe and predict the behavior of the phenomenon. |
| **Empirical Support** | A theory is grounded in empirical evidence. It should be supported by a substantial body of observations, experiments, and data collected through systematic and repeatable methods. |
| **Evolution** | Scientific theories are subject to revision as new evidence and understanding emerge. |
| **Consistency** | A scientific theory must be internally consistent, meaning its various components and principles should not contradict each other. |

Essential elements for a Theory of Awareness

# Domain

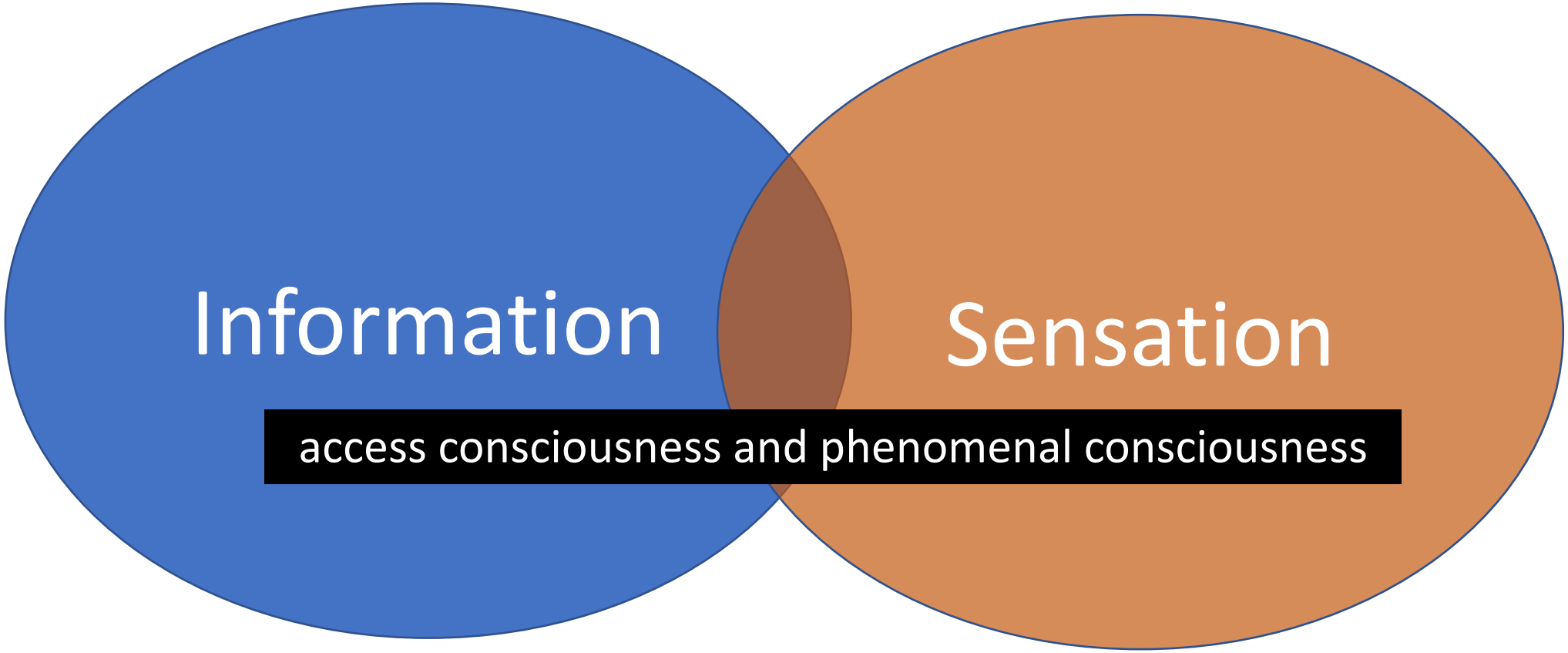
A theory explains a specific natural phenomenon
or a particular domain of inquiry.
It defines the scope of what it seeks to explain.

# Two (related) Terms

Awareness

Consciousness

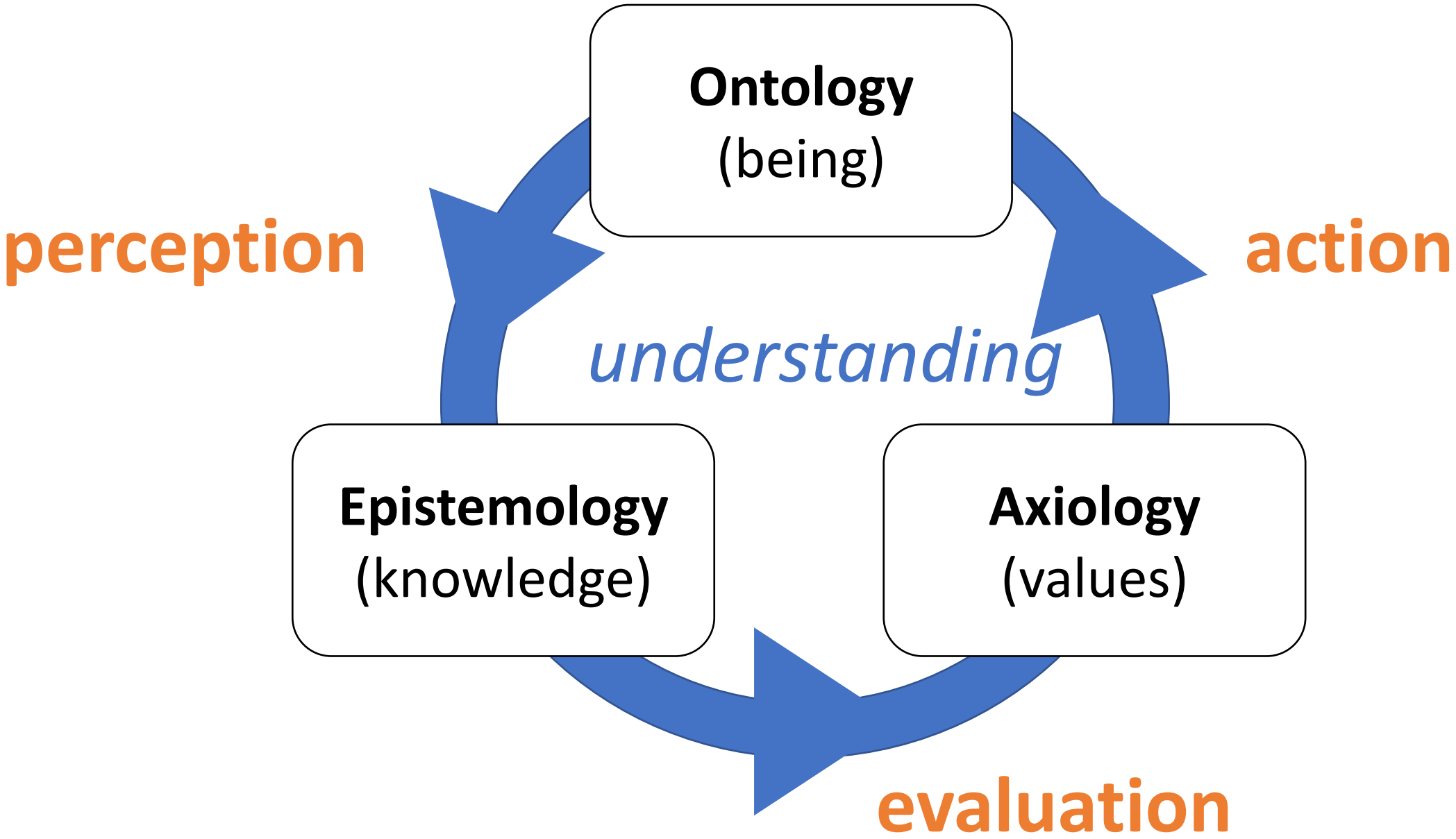
# Two (related) Aspects




N. Block, “On a confusion about the function of consciousness,” Behav. Brain Sci., vol. 18, pp. 227–247, 1995.

# Four (philosophical) targets for awareness

| What is there | What is felt |
| --- | --- |
| Ontology | Phenomenology |
| Epistemology | Axiology |
| What is known | What is important |


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

Essential elements for a Theory of Awareness

# Concepts

A theory contains a set of well-defined concepts that provide the vocabulary and framework for discussing and understanding the phenomenon.

# Essential concepts for a formal ToA

***Sensing***: The production of information for the subject from an object.

***Perceiving***: The integration of the sensory information bound to an object into a model of the object.

***Model***: Integrated actionable representation; an information structure that sustains a modelling relation.

***Engine***: Set of operations over a model.

***Inference***: Derive conclusions from the model. Apply engines to model.

***Valid inference***: A inference whose result matches the phenomenon at the modelled object.

***Exert a model***: Perform valid inferences from the model.

***Understanding***: Achieving exertability of a model of the object/enviroment.

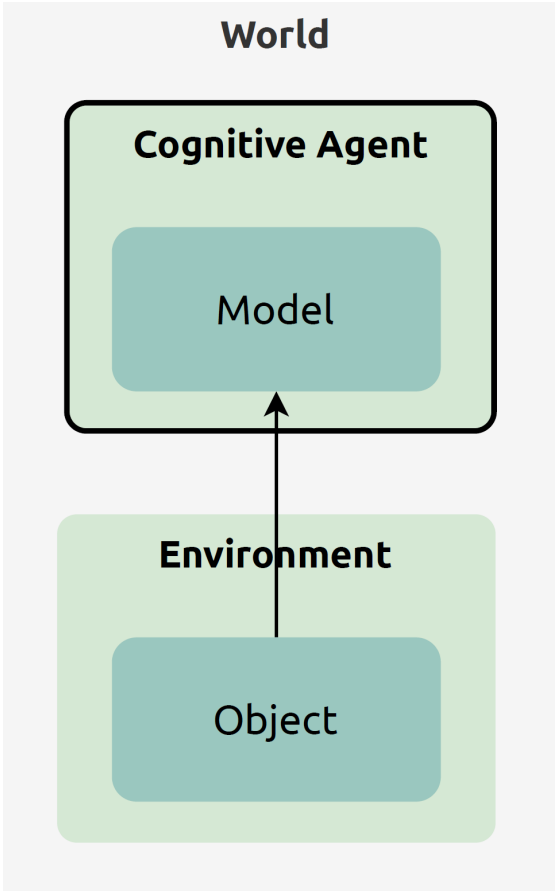
***Specific understanding***: Understanding concerning a specific set of exertions.

***Mission understanding***: Understanding concerning a set of mission-bound exertions.

***Omega understanding***: Understanding all possible exertions of a model in relation to an object.

***Awareness***: Real-time understanding of sensory flows.

***Self-awareness***: Subject-bound awareness. Awareness concerning inner perception.



***Awareness***:

Understanding of sensory flows.

***Understanding***:

Having an actionable model of the object/environment.

# Towards a Theory of Awareness

**Ricardo.Sanz@upm.es**

CORESENSE