Towards a Theory of Awareness

Ricardo Sanz[1][a], Manuel Rodríguez[1][b], Martín Molina[2][c],
Esther Aguado[1][d], and Virgilio Gómez [1][e]
[1]*Autonomous Systems Laboratory, Universidad Politécnica de Madrid,*
c/ José Gutierrez. Abascal 2, 28006 Madrid, Spain
[2]*Artficial Intelligence Department, Universidad Politécnica de Madrid,*
Campus de Montegancedo, 28660 Boadilla del Monte, Madrid, Spain
{*ricardo.sanz, manuel.rodriguezh, martin.molina, e.aguado, virgilio.gomez.lambo* }*@upm.es*

Keywords: Awareness, Theory, Autonomy, Cognitive Science, System Architecture, Engineering Methods, Science, Consciousness, Perception, Attention, Understanding.

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 concering 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*.

1 INTRODUCTION

The mere observation of humans' and animals' behaviours shows that *awareness* is a central aspect of mentality for those agents that successfully operate in changing, challenging environments. Perceiving change, responding to potentially dangerous situations and leveraging environmental affordances are essential capabilities for autonomous agents ecological viability, i.e. agents that are capable of surviving the disturbances that a non-controlled environment throw on them. So should be for robots and other classes of situated machines. Endowing machines with "awareness" *should improve their autonomy profile, making them more reactive to world dynamics* (Sanz et al., 2007a). However, this is not an easy task; especially because we do not have a good, translatable **Theory of Awareness** (ToA).

In this domain, authors sometimes use the term "awareness" and sometimes use the term "consciousness". The use of the two terms has similarities and differences that may vary across disciplines, domains and languages. It is not easy to pinpoint the difference between "consciousness" and "awareness", as Francis Crick acknowledged (Crick and Koch, 1992). From some perspectives, we can consider both the same thing, e.g. in the perceptual domain; from others, we cannot, e.g. in the ethical domain. We will follow here Crick's policy of considering them synonyms; giving preference to "awareness" and using "consciousness" only when trying to make a specific point concerning the term. However, other authors do follow other policies and use both terms (as the references and theories will show).

In principle, it should be feasible to translate most human mental traits into machine mental traits thanks to the general multiple realizability of physical sys-

[a] https://orcid.org/0000-0002-2381-933X
[b] https://orcid.org/0000-0003-0929-5477
[c] https://orcid.org/0000-0001-7145-1974
[d] https://orcid.org/0000-0002-7860-9030
[e] https://orcid.org/0000-0001-8538-5111

tems[1]. To translate awareness from humans to machines we only need a *physical* theory of awareness. However, in the case of *awareness-related aspects*, there is a strong risk of falling into a very deep rabbit hole of biologism if we are not careful enough as to be perfectly clear on what we are talking about and what class of theory do we need. In this paper we *argue for the development of such a general theory of awareness*, trying to be careful enough as to avoid the rabbit hole, and delivering a theory applicable to animals and translatable to machines.

2 AWARENESS IN ROBOTS

The topics of consciousness and awareness have been quite marginal in the domains of AI and robotics (Chella and Manzotti, 2007; Chella, 2023).

The efforts to build conscious AIs or the discussion about its very possibility was a matter of a small group of researchers with some philosophical vocation. Only recently this community has gained some extra human mass, esp. when some comments by well-known actors in the current AI landscape hit the media. These comments talked about the possibility of current mainstream AI implementations could be reaching a state of consciousness. These reminded people about Skynet becoming aware and trying to kill humanity in Cameron's classic *Terminator*. However, even when this is becoming a wider discussion, the research program on aware AI is still very flaky; but it should be not if we are right concerning its importance for autonomous machines.

In the last decades, some researchers have attempted the creation of real implementations of AIs and robots with consciousness. In most cases, the approach take was 1) select one of more characteristics associated to human consciousness, and 2) develop a machine that demonstrated this characteristic. For example, Aleksander addressed anticipation and imagination (Aleksander, 2009), Tani addressed mirror self recognition (Tani, 2017), Chella addressed inner speech (Chella et al., 2020), Hoffmann addressed proprioceptive self-awareness (Hoffmann, 2021), Hernández addressed metacognitive self-awareness (Hernández et al., 2009), and, brave enough, Haikonen addressed qualia (Haikonen, 2013).

The attempts will continue, for example extending into the very active domains of large, language-based AIs. However, the current statistical approaches of machine learning from human-bound data will not achieve the desired end of powering the engineering capability of creating *custom awareness*[2]. The engineering of these systems will only be effective if grounded in more profound, structural theories that seem far from what current machine learning capability can provide. As Dacey says, (Dacey, 2022) "Statistical inference cannot do all of the work of theory choice." We have the need of producing a proper structural deep theory of awareness to do both science and technology.

3 RATIONALE FOR A THEORY OF AWARENESS

The rationale for seeking a theory of awareness has deep grounds in science, philosophy and engineering (Sanz et al., 2007a).

Understanding awareness needs an approach from a unified perspective. For example, in the realm of philosophy, awareness is related to four essential branches: epistemology -what the agent gets to know through the senses, ontology -what the agent is aware of-, phenomenology -the sensations that the agent gets-, axiology -the value that such perceptions have for the agent dwellings. Awareness seems essential for dwelling in open dynamic worlds.

This need is also very relevant in the world of artificial systems. Engineers are seeking system design solutions to deal with the uncertainty of the world and the uncertainty of the systems themselves. In many cases, autonomous robots failures are not due to changes in the world but changes in the robot itself (e.g. failures or emergent phenomena in the robot software subsystems). The streamlined construction and dependable runtime operation of artificial autonomous systems depend on using suitable system architectures and applying proven engineering methods (Aguado et al., 2021). Having solid theories of world-awareness and self-awareness may help achieve the desired results.

In this sense, a general theory of awareness is a desirable asset for both scientists and engineers. Unfortunately, the scientific analysis of the awareness phenomena is quite difficult because it cannot be easily identified and isolated in fully functioning biological minds. We need a single theory of general awareness of universal nature and this universality seems difficult, especially when we seek a theory applicable both to humans and robots, and not exclusively bound

[1] We will here take a strict, non-dualist, physicalist stance concerning mental aspects like awareness.

[2] Awareness that is designed and scaled to the needs of the target technical system that need not be similar to a human.

to the psychology and neurobiology of living animals (Wilson, 1998). Achieving an *universal* –applicable to all classes of systems– and *unified* –explaining all related phenomena— theory is a complex challenge. Besides the different classes of subjects, there are too many aspects of consciousness to deal with. Aleksander identified five aspects of consciousness to be mapped into machines —perception, imagination, attention, planning, emotion— but, for example, Tani identified ten different aspects —based on the phenomenological analysis of Husserl.

Long ago, Aaron Sloman said that "It is not worth asking how to define consciousness, how to explain it, how it evolved, what its function is, etc., because there's no one thing for which all the answers would be the same. Instead, we have many sub-capabilities, for which the answers are different: e.g. different kinds of perception, learning, knowledge, attention control, self-monitoring, self-control, etc."[3]

Many authors, especially in the biological and humanities domains, argue that machine consciousness is impossible. However, the possibility of devising a single mechanism explaining all the aspects of awareness in a system-neutral sense, or at least separating them into related and unrelated aspect as Sloman proposed, need not be an impossible dream (Hadley, 2023). As Francis Crick said:

> The second assumption is tentative: that all the different aspects of consciousness, for example pain and visual awareness, employ a basic common mechanism or perhaps a few such mechanisms. If we understand the mechanisms for one aspect, we will have gone most of the way to understanding them all. (Crick and Koch, 1990)

In this position paper, the authors argue for deploying effort towards developing such a theory, stating some requirements for it and proposing an initial seed for a potential future *Theory of Awareness*.

4 THEORIES OF AWARENESS

There are many theories that offer different perspectives on the nature of awareness, reflecting the interdisciplinary nature of the field. See for example the article by Seth and Bayne (Seth and Bayne, 2022) for a more exhaustive review in the domain of biological consciousness.

4.1 Perspectives on Awareness

Here are some important theories and perspectives related to awareness:

- In the domain of cognitive psychology:
  - Selective Attention: Awareness seems closely related to selective attention. We are consciously aware of the information we selectively attend to, and other stimuli may not enter our conscious awareness (Taylor, 2002).
  - Levels of Processing: Depth of processing information affects awareness (Craik and Lockhart, 1972). Deeper elaboration leads to better retention and awareness of information.
- In the domain of neuroscience:
  - Global Workspace Theory (GWT): GWT (Baars, 1997) says that conscious awareness arises from the global broadcast of information throughout the brain. Certain neural processes involve a global workspace that integrates information and brings it to conscious awareness.
  - Neural Correlates of Consciousness (NCC): Researchers seek to identify specific neural activity patterns associated with conscious awareness (Koch et al., 2016). Understanding these neural correlates can shed light on how the brain generates awareness.
- In the domain of Philosophy:
  - Phenomenal Consciousness: This perspective explores the nature of subjective experience — having "qualia." It analyses what it is like to have a particular experience and how subjective awareness happens (Nagel, 1974).
  - Higher-Order Thought (HOT) Theory: This theory proposes that awareness arises from the ability to have thoughts about one's own mental states (Rosenthal, 2000).
- In the domain of social psychology:
  - Social Awareness: How individuals are aware of and respond to the social environment (Durkheim, 1893). It includes theories related to social perception, empathy, and understanding the mental states of others.
- In the domain of general systems:
  - Integrated Information Theory (IIT): IIT (Tononi, 2004) says that a system is conscious if it has a high degree of integrated information. This theory has received continuous support thanks to its formal nature.
- In the domain of quantum mechanics:

[3] In comp.ai.philosophy, 14 Dec. 1994.

- Orchestrated Objective Reduction (Orch-OR): Orch-OR (Hameroff and Penrose, 2014) suggests that quantum processes in microtubules within brain cells play a role in consciousness.

Besides the effort put and the raising interest in the topic, the study of awareness is still in a pre-scientific stage. There are lots of ongoing research and debate, but no single theory has gained unanimous acceptance (Jylkkä and Railo, 2019). See, for example, the recent polemics about IIT being considered as pseudoscience by some authors, esp. in the neuropsychology domain (Fleming, 2023).

4.2 Awareness vs Consciousness

As said at the beginning, the terms "awareness" and "consciousness" are often used interchangeably. However the two terms are sometimes used differently in various contexts, and there isn't a universally agreed-upon distinction.

Different perspectives —GWT, IIT, Orch-OR, etc.— attempt to differentiate awareness from consciousness but these distinctions are not universally accepted. Different theories and disciplines may use the terms in varying ways and explain them using conceptualizations that are far from beign harmonised. The field remains dynamic, and our understanding of consciousness and awareness is likely to evolve with ongoing research and interdisciplinary exploration and the elaboration of proper, transversal theories.

A distinction that may be somewhat distilled from the previous list is that awareness is related to information –i.e. to the epistemic aspect– and consciousness to sentience –i.e. to the phenomenic aspect. The work that we are doing points into the epistemic direction, leaving the sentience aspect to ulterior scientific efforts.

5 THE CONTENT OF A THEORY

The ToA shall be a theory that is both scientific and operationalisable. As a scientific theory, the ToA shall be a well-substantiated explanation of some aspect of the natural world that is based on a body of evidence, observations, and experiments. This body of evidence comes from the cognitive operation of animals and also from the cognitive operation of machines, esp. in the uncertain domains of the open robot world. The "aspect of the natural world" that we are interested in is the phenomenon of "awareness".

We want the ToA to be a solid theory —a robust and well-established scientific explanation— to serve as the framework upon which scientists could build their understanding of the phenomenon of awareness and pile-up solid research results, and engineers use this understanding in its operationalisation as applied science in building better cognitive robots.

5.1 Characteristics of a Scientific Theory

This is a list of key components and characteristics of a scientific theory and to what extent the ToA shall address them:

Empirical Basis: A scientific 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. This empirical support is crucial in distinguishing a theory from a mere hypothesis or conjecture. The source of evidence is deployed human cognition and the situations where awareness plays a central role. For example, the METATOOL project[4] explores evidence concerning the role of awareness in ancient tool making. Obviously we are dealing with cognitively problematic situations (Norman, 1980) and hence, a proper organization of evidence will be critical.

Consistency: A scientific theory must be internally consistent, meaning its various components and principles should not contradict one another. It should provide a logical and coherent framework for explaining observed phenomena. It shall also be consistent with other accepted scientific theories. The use of formal methods –as IIT attempted– and the model-based methods of engineering may provide the necessary support to guarantee this consistence.

Testability: Scientific theories are falsifiable (Popper, 1959). They can be subjected to experimentation and observation, and there should be clear criteria that, if not met, would disprove the theory. The ability to test and potentially disprove a theory is a fundamental aspect of the scientific method. The level of robot performance in real settings may provide this necessary evidence.

Predictive Power: A strong scientific theory can make predictions about future observations or experiments. These predictions should be based on the theory's principles and should be verifiable through empirical testing. The theory's ability to make accurate predictions lends further credibility to it. This is an essential aspect for a theory that is

[4] http://metatool-project.eu

used as a design asset in an engineering endeavour. Engineers will use designs based on this theory to guarantee effectiveness in future systems.

Scope and Explanatory Power: A scientific theory should have a broad scope, meaning it can explain a wide range of related phenomena. The more phenomena it can explain, the more powerful and influential the theory is. This sits at the very core of the scientist ambition: we target a theory that not only addresses *robotic awareness* but *general cognitive systems awareness* (see Figure 1).

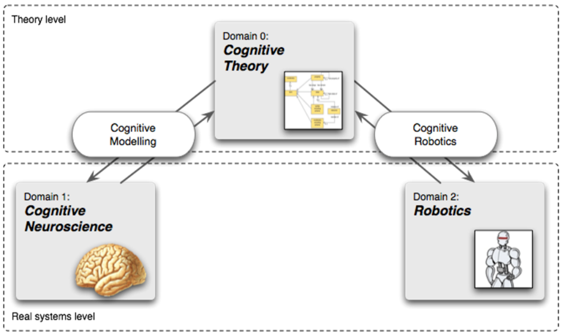

Figure 1: Target a general cognitive theories of awareness.

Simplicity: If two or more theories explain the same phenomena equally well, the simpler one is to be preferred (Ockam's razor). Simplicity makes theories more elegant and easier to work with. So far, there is no real competition in theories of awareness (*i.e.* in the terms stated here). We shall try to make the ToA simple using compact and effective abstractions. The CORESENSE project[5] tries to use category theory for these abstractions in part motivated by this search for simplicity.

Reproducibility: The experiments should be reproducible by other scientists using the same methods and conditions. This is a cornerstone of the scientific method and ensures the reliability of the theory. This has always been a problem in cognitive robotics and in cognitive systems in general. To this end, benchmarking has been used in robotics to enhance this reproducibility. For example, the RoboCup@Home challenge used in robotics specifically addresses this problem.

Peer Review: Experts in the field assess the theory's validity and the quality of the evidence supporting it. Peer review is an important process for maintaining the rigor and reliability of scientific theories. The Open Science approach of modern research specifically addresses this need.

Consensus : While scientific consensus can evolve and change, a well-established theory is generally widely accepted within the scientific community (achieving the status of "normal science" in Kuhn terms or becoming a "framework of shared commitments" (Eckardt, 1995)). In the domain of psychology this is not easy. Psychological constructs suffer from the "toothbrush" problem: no self-respecting psychologist wants to use anyone else's. But in rigorous science, we are sometimes forced to do so by the force of the facts. If our theory is solid, well documented and effective, consensus will eventually emerge.

Understandability : As quantum mechanics demonstrate, understandability is not a necessary characteristic of scientific theory. However, we would like our theory to be understandable by a broad community of scientists and engineers. This may require from us the expression of the theory in different ways that can reach these people.

In science, alive theories are not absolute truths but are just our best current explanations based on the available evidence. The process of forming and refining scientific theories is the essential ongoing and dynamic aspect of the scientific endeavour. In a sense, this filtering-out by evidence can only happen when the abstract concepts are strictly mapped into more concrete realities. Most of the discussion on awareness is pre-scientific in this sense: deals with abstractions disconnected from the evidence.

5.2 Operationalisation of a Scientific Theory

The term "operationalizable theory" is not a standard concept in scientific terminology. It is a combination of two important concepts: "theory" and "operationalization."

Theory: As described before, a theory is a well-substantiated explanation of some aspect of the natural world that is based on empirical evidence and can be used to make predictions and understand phenomena.

Operationalization: Operationalization is a process in research where abstract concepts or variables are defined and measured in a concrete and observable way. It involves specifying how a theoretical concept will be measured or observed in practice. This is a crucial step in turning abstract theories into testable hypotheses and conducting empirical research.

The term "operationalizable theory" describes a theory that has been sufficiently developed and de-

[5] http://coresense.eu

fined so that its key concepts and variables can be operationalized for empirical experimentation. In this context, an operationalizable theory would be one that can be translated into specific, measurable variables or constructs that researchers can work with to conduct experiments, gather data, and test hypotheses (e.g. building AI-driven robots and deploying them in open worlds). This operationalization is a crucial step in the scientific method when examining the applicability and validity of a theory in real-world scenarios.

6 INITIAL STEPS TO A THEORY OF AWARENESS

In the CORESENSE and METATOOL projects we have the specific task of developing a ToA. Any theory is a complex construct that consists of several key elements. These elements shall work together to provide a comprehensive and well-substantiated view of some aspect of the worlds. In some sense this view is *explanatory* —*e.g.* when applied to humans or animals— and in another sense this view is *operational* —as when it enables the construction of artefacts.

6.1 Essential Elements for a Theory

Table 1 describes some specific elements and aspects that a theory typically has. It is too early in the development of our ToA to detail how the theory addresses all of them. In this paper we will just address the **domain** and the initial **concepts**. The domain is, obviously, "awareness". The initial concepts are summarily described in Section 6.2.

The elements and structure of a theory may vary depending on the field of science and the nature of the phenomenon being studied. This may imply some specifics of the ToA when applied to certain classes of systems. However, these elements collectively contribute to the development of a robust and well-supported scientific theory.

6.2 Initial concepts for a Theory of Awareness

When considering the domain of the theory, we shall be aware that both consciousness and awareness are mongrel concepts when applied to humans: They are used in many senses, referring to different classes of phenomena, generating confusion and long irrelevant discussions.

An analytical effort is necessary to separate the different mental aspects they refer to and a terminological effort will be necessary to suitably label all those aspects. In some cases, the terminological effort is addressed by using noun phrases like "visual awareness" or "synthetic awareness", but this usually implies a subclassing from a general "awareness" class. Another approach is the use of prefixes to create new terms when the intention is to create terms for unrelated classes of phenomena. Examples of this are P-consciousness and A-consciousness for phenomenal consciousness and access consciousness.

In this paper the terms are used in a very specific sense: we are interested in the *functional aspects of awareness* that let an **autonomous system** act properly when it is aware of the situation it is immersed in. In our case, the situational awareness proposed by Endsley (Endsley, 1995) address not only the system environment but extends to 1) the system itself and 2) its relation to the environment when pursuing an externally imposed mission. The fundamental domain entities in this scenario are: the autonomous system is the **subject** of awareness, generating a mental **model** of some **object**s situated in the **environment**, the part of the **world** that is causally connected with both subject and objects.

In this base scenario of a subject being aware of an object, we investigate the essential character of awareness processes and how they are related to perception and understanding. We are interested in the processes that underlie the capability of the subject to be aware and understand the changes in the object to be more effective in completing its mission.

As part of these initial steps, these are some fundamental concepts under elaboration in this research:

***Sensing*:** Getting information –sense-data– bound to an object in the environment.

***Perceiving*:** Integration of the sensory information into a model of the object by means of a modelet.

***Modelet*:** A partial model related to a target system. A information structure that sustains a modelling relation (Rosen, 2012).

***Model*:** Integrated actionable representation; an integrated set of modelets.

***Engine*:** Set of operations over a model (e.g., integration of a modelet, exertion, compaction, pruning, intensification, chunking, etc).

***Inference*:** Derive conclusions from the model.

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

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

Table 1: Essential Elements and Aspects of a Theory.

Element	Content of the element
Phenomenon or 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, i.e. testable predictions about how certain variables or factors are related within the defined domain, and guide empirical research.
Laws or Principles	A theory may incorporate laws or principles typically derived from empirical data and observations, that describe relationships or patterns observed within the phenomenon.
Relationships	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 behaviour 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.
Reviews	Scientific theories are typically subjected to peer review, where experts in the field assess the theory's validity and the quality of the evidence supporting it.

Understanding**:** Achieving exertability of a modelet; e.g. by mental model integration of a modelet and activation of associated engines.

Specific understanding**:** Understanding concerning a specific set of exertions (can be extensive or intensive).

Structure understanding**:** Understanding the structure of the object implies achieving exertability concerning the system structure.

Behaviour understanding**:** Understanding the behaviour of the object (achieving exertability concerning the system behaviour).

Mission understanding**:** Understanding mission-bound exertions (i.e. achieving derivability of valid results from the model that can be used by the agent to fulfill the mission).

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

Awareness of**:** Object-bound awareness.

Self-awareness**:** Subject-bound awareness. Awareness concerning inner perceptive flows.

The current project work is related to the development of the formal model of the theory of awareness in the form of 1) a formal ontology in higher-order logic and 2) an architecture expressed in formal MBSE[6] languages.

7 CONCLUSIONS

The elaboration of formal theories of awareness is active (Bringsjord and Sundar, 2020) but still lost in the variety of phenomena associated to consciousness. An analytical approach at clarifying the many aspects of mentality and awareness is necessary. This means properly setting the boundaries of *the domain of explanation* and being more precise on the class of phenomena that the theory is addressing.

In this paper we have described the overall expected content of a Theory of Awareness to be of applicability in the construction of broad classes of autonomous systems. The elaboration of the causal principles and laws will enable the development of architectural patterns that will enable the design of new systems (Sanz et al., 2007b). Besides architectural patterns, this theory could also provide many other resources in autonomous systems engineering to analyze existing systems, to guide new designs, or to build specific programs, etc. (examples of specific re-

[6] MBSE is an acronym of Model-Based Systems Engineering, a way of performing systems engineering where models are the central assets.

sources are design idioms, complete reference architectures, common terminologies formalized through ontologies, domain specific languages, reusable components in software libraries, etc.).

See more about these developments at the CORESENSE[7] and METATOOL[8] project websites and the Awareness Inside[9] EIC Pathfinder Challenge.

ACKNOWLEDGMENTS

The authors acknowledge the support of the European Commission through the Horizon Europe project #101070254 CORESENSE and the European Innovation Council project #101070940 METATOOL.

REFERENCES

Aguado, E., Milosevic, Z., Hernández, C., Sanz, R., Garzon, M., Bozhinoski, D., and Rossi, C. (2021). Functional self-awareness and metacontrol for underwater robot autonomy. *Sensors*, 21(4):1–28.

Aleksander, I. (2009). Designing Conscious Systems. *Cognitive Computation*, 1(1):22–28.

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

Bringsjord, S. and Sundar, G. (2020). The Theory of Cognitive Consciousness, and Λ (Lambda). *Journal of Artificial Intelligence and Consciousness*, 07:1–27.

Chella, A., editor (2023). *Computational Approaches To Conscious Artificial Intelligence*. World Scientific.

Chella, A. and Manzotti, R., editors (2007). *Artificial Consciousness*. Imprint Academic.

Chella, A., Pipitone, A., Morin, A., and Racy, F. (2020). Developing Self-Awareness in Robots via Inner Speech. *Frontiers in Robotics and AI*, 7:16.

Craik, F. I. and Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, 11(6):671–684.

Crick, F. and Koch, C. (1992). The Problem of Consciousness. *Scientific American*, 267(3):152–159.

Crick, F. and Koch, C. C. (1990). Towards a neurobiological theory of consciousness. *Seminars in the Neurosciences*, 2:263–275.

Dacey, M. (2022). Separate substantive from statistical hypotheses and treat them differently. *Behavioral and Brain Sciences*, 45:e9.

Durkheim, É. (1893). *De la division du travail social*. Presses Universitaires de France.

Eckardt, B. V. (1995). *What is Cognitive Science?* Bradford Books. The MIT Press.

Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors*, 37(1):32–64.

Fleming, S. M. e. a. (2023). The Integration Information Theory As Pseudoscience.

Hadley, M. J. (2023). A Generic Model of Consciousness. *Journal of Artificial Intelligence and Consciousness*, 10(02):291–308.

Haikonen, P. O. A. (2013). Consciousness and the Quest for Sentient Robots. In Chella, A., Pirrone, R., Sorbello, R., and Jóhannsdóttir, K., editors, *Biologically Inspired Cognitive Architectures 2012*, pages 19–27. Springer.

Hameroff, S. and Penrose, R. (2014). Consciousness in the universe. *Physics of Life Reviews*, 11(1):39–78.

Hernández, C., López, I., and Sanz, R. (2009). The Operative Mind: a Functional, Computational and Modeling Approach To Machine Consciousness. *International Journal of Machine Consciousness*, 1(1):83–96.

Hoffmann, M. e. a. (2021). Robot in the Mirror: Toward an Embodied Computational Model of Mirror Self-Recognition. *KI - Kunstliche Intelligenz*, 35(1):37–51.

Jylkkä, J. and Railo, H. (2019). Consciousness as a concrete physical phenomenon. *Consciousness and Cognition*, 74:102779.

Koch, C., Massimini, M., Boly, M., and Tononi, G. (2016). Neural correlates of consciousness: progress and problems. *Nat Rev Neurosci*, 17(5):307–321.

Nagel, T. (1974). What Is It Like to Be a Bat? *The Philosophical Review*, 83(4):435–450.

Norman, D. A. (1980). Twelve Issues for Cognitive Science. *Cognitive Science*, 4:1–32.

Popper, K. (1959). *The Logic of Scientific Discovery*. Basic Books.

Rosen, R. (2012). *Anticipatory Systems. Philosophical, Mathematical, and Methodological Foundations*, volume 1 of *IFSR International Series on Systems Science and Engineering*. Springer, 2nd edition.

Rosenthal, D. M. (2000). Metacognition and Higher-Order Thoughts. *Consciousness and Cognition*, 9:231–242.

Sanz, R., López, I., and Bermejo-Alonso, J. (2007a). A Rationale and Vision for Machine Consciousness in Complex Controllers. In Chella, A. and Manzotti, R., editors, *Artificial Consciousness*. Imprint Academic.

Sanz, R., López, I., Rodríguez, M., and Hernández, C. (2007b). Principles for consciousness in integrated cognitive control. *Neural Networks*, 20(9):938–946.

Seth, A. K. and Bayne, T. (2022). Theories of consciousness. *Nature Reviews Neuroscience*, 23(7).

Tani, J. (2017). *Exploring robotic minds : actions, symbols, and consciousness as self-organizing dynamic phenomena*. Oxford University Press.

Taylor, J. G. (2002). How Can the Self Understand Itself? A Review of Models of the Self Edited by Shaun Gallagher and Jonathan Shear. *Psyche*, 8(12).

Tononi, G. (2004). An information integration theory of consciousness. *BMC Neuroscience*, 5(42).

Wilson, E. O. (1998). *Consilience. The Unity of Knowledge*. Alfred A. Knopf, New York, USA.

[7] http://coresense.eu

[8] http://metatool-project.eu

[9] https://awarenessinside.eu/