

# Towards a Definition of Awareness for Embodied AI

Giulio Antonio Abbo<sup>1</sup><sup>a</sup>, Serena Marchesi<sup>2</sup><sup>b</sup>, Kinga Ciupinska<sup>2</sup><sup>c</sup>, Agnieszka Wykowska<sup>2</sup><sup>d</sup>  
and Tony Belpaeme<sup>1</sup><sup>e</sup>

<sup>1</sup>IDLab-AIRO, Ghent University, imec, Belgium

<sup>2</sup>Social Cognition in Human-Robot Interaction (S4HRI), Italian Institute of Technology, Genova, Italy

Keywords: Awareness, Artificial Intelligence, Embodied AI.

Abstract: This paper explores the concept of awareness in the context of embodied artificial intelligence (AI), aiming to provide a practical definition and understanding of this multifaceted term. Acknowledging the diverse interpretations of awareness in various disciplines, the paper focuses specifically on the application of awareness in embodied AI systems. We introduce six foundational elements as essential building blocks for an aware embodied AI. These elements include access to information, information integration, attention, coherence, explainability, and action. The interconnected and interdependent nature of these building blocks is emphasised, forming a minimal base for constructing AI systems with heightened awareness. The paper aims to spark a dialogue within the research community, inviting diverse perspectives to contribute to the evolving discipline of awareness in embodied AI. The proposed insights provide a starting point for further empirical studies and validations in real-world AI applications.

## 1 INTRODUCTION

The concept of *awareness* does not allow itself to be pinned down easily. Indeed, the term is found across virtually any discipline, not rarely with different meanings. Often it is used to describe voluntarily directing one's attention towards a certain aspect. In other cases, it has a specific and circumscribed meaning, seldom familiar to the uninitiated outside a particular field.


In philosophy, awareness pertains to consciousness and self-awareness, with philosophers investigating how mental states interconnect with physical processes in the mind-body problem (Fodor, 1981). Psychology finds it closely linked to consciousness and delves into different levels of awareness, ranging from the conscious to the subconscious and unconscious. Neuroscience sheds light on the neural correlates of awareness, studying brain activity associated with conscious experiences. Researchers in neuroscience also explore altered states of consciousness,


such as sleep, meditation, and drug-induced states, to unravel the neural mechanisms underlying awareness.


In this turmoil of different definitions, a complete reconciliation is unfeasible. Furthermore, each of these definitions is dictated by the heterogeneity of their applications. Thus, an unifying attempt would be counterproductive and limiting, as it would lose the necessary specificity and detail.


We will focus on awareness applied to the field of artificial intelligence (AI). In particular, we will discuss what awareness means when dealing with embodied AI (Chrisley, 2003; Pfeifer and Bongard, 2006; Duan et al., 2022).


Using its body an AI system can explore its surroundings using sensorimotor behaviour, implying that embodied AI has a certain level of control – or *agency* – over what it does in the environment. For example, a social robot can use its camera feed to interpret visual events near it and respond appropriately, and in a multi-party conversation, the robot can use a microphone array to distinguish between speakers and provide insights into what was discussed. The applications for embodied AI are countless and we expect embodied AI to acquire even greater relevance in our everyday lives, thanks to the advent of Large Language Models (LLM) and specifically Multimodal LLMs.

<sup>a</sup> <https://orcid.org/0000-0001-6301-0028>

<sup>b</sup> <https://orcid.org/0000-0001-9931-156X>

<sup>c</sup> <https://orcid.org/0000-0002-9909-4400>

<sup>d</sup> <https://orcid.org/0000-0003-3323-7357>

<sup>e</sup> <https://orcid.org/0000-0001-5207-7745>

Given the ever-changing dynamics of the social and physical world, having access to sensor data is not sufficient. Indeed, it is necessary to integrate the information, enabling coherent interactions between internal representation, and eventually the system and the outside world. As a consequence of their increased autonomy from human intervention, these algorithms will necessarily have to show an advanced level of something which might best be described as “awareness”. A quality which allows a system to exhibit optimal performance, enabling them to interact efficiently with their social and physical environment and respond contingently and quickly to dynamic situations.

Consider for instance a system operating in a social environment, where interaction with humans is fundamental. This can be the check-in area of a busy metropolitan hospital, with all the complexities associated with it: from the noisy environment to the wide range of cultures and ages. The system could take the form of a robot in the future, but let us imagine for now an interactive kiosk. Currently, where kiosks are available, they show an interface that hustles the patients through the initial part of the check-in procedure. This experience is often not pleasant, especially the first time around, and would greatly benefit from a more human-centric approach. The system could be empowered with an AI to facilitate a smoother and more user-friendly check-in process, for example by being aware of the patient’s emotions. This goes beyond the simple detection of a smile, as it involves establishing a common ground and complex topics such as the Theory of Mind (Frith and Frith, 2005). The result would be an enhanced and more humanised form of interaction, which is fundamental in delicate scenarios, such as the one presented.

In this paper, we identify and discuss six building blocks of an aware embodied AI, showing how each of them is necessary and providing an illustrative application. Then, we suggest how the items presented are connected and interdependent, before moving to propose a definition of awareness in this field. While our definition is a working definition, this paper contributes to the discourse on awareness and consciousness in AI by offering new thoughts and perspectives, thereby enriching the ongoing exploration in this field.

## 2 BACKGROUND

### 2.1 Consiusness, Awareness, Self-Awareness

Dehaene et al. (Dehaene et al., 2017) identify two levels of consciousness: *Global Availability* and *Self-Monitoring*. The first represents consciousness in its transitive meaning as in *being conscious of X*, the information becomes globally available to the rest of the system for further processing. Fundamental at this level is the mental representation of the object of thought and the ability to report about it verbally and non-verbally. The psychological definition of *attention* (James, 1890) overlaps with the concept of global availability if we exclude the previous stages of involuntary attentional selection. The second level identifies the reflexive meaning of consciousness, in the sense of self-monitoring, introspection, or meta-cognition. Confidence, reflection, meta-memory and reality monitoring are all aspects related to this level of consciousness or self-awareness. Importantly, the two levels are orthogonal as one can exist without the other.

In their position paper, Dehaene et al. state that a machine endowed with these two capabilities would behave as if conscious. However, a holistic implementation in which the system becomes aware of everything is currently technologically unfeasible. For a concrete application of the definition, we find it necessary to always specify the object of the awareness. We will thus refer to awareness *of* something: e.g., awareness of the self (self-aware), awareness of the context, awareness of our capabilities, and so on. This limitation, in which a system is aware only of a few aspects, carefully avoids any conscious-mimicking behaviour.

### 2.2 Embodied AI

“Embodied AI is about incorporating traditional intelligence concepts from vision, language, and reasoning into an artificial embodiment” (Duan et al., 2022). Conventional AI leverages the vast amounts of data available on the internet from text, to multimedia elements, to the most diverse datasets. On the other hand, embodied AI integrates physical interaction and sensorimotor capabilities into artificial agents. The physical presence of embodied AI needs to be reflected in its training data. For this reason, egocentric (first-person) perception plays a central role in this field. First-person data consists of videos and images taken from the point of view of the agent, in this case, the embodied system. With this new kind of

data, it is possible to tackle new and exciting problems (Grauman et al., 2022): indexing past experiences, analysing present interactions, and anticipating future activity.

Chella et al. set to achieve awareness – specifically, self-awareness – through inner speech (Chella et al., 2020). Inner speech can take many forms: it can consist of just a few words or full sentences, and it can be a monologue or a dialogue, in the case one asks questions and answers them using both “I” and “You”. This process is involved in self-regulation, language functions such as writing and reading, remembering the goals of action, task-switching performances, Theory of Mind, and self-awareness. The system makes use of perception and action modules, it includes proprioception of emotions, beliefs, desires, intentions and body as well as exteroception. Actuators include covert articulation and motor modules, and everything is enabled by a set of memory modules (Chella and Pipitone, 2020). We choose to focus on the indispensable ingredients of awareness, since we are not set to achieve true consciousness. This means that several aspects, such as beliefs and desires – but also goals – are not considered in our work.

### 3 ELEMENTS FOR AWARENESS

What are the building blocks for an aware embodied AI? Which processes, structures and properties are required for an aware behaviour when dealing with the external world? In this section, we introduce six requisites: access to information, information integration, attention, coherence, explainability, and action. We purposefully will not cover those aspects that are secondary nice-to-haves but do not constitute a minimal base for awareness.

#### 3.1 Access to Information

For a system to be aware of  $X$ , it must have access to  $X$ . While this statement may seem self-evident, we want to stress that access, in this context, extends beyond mere availability. It encompasses the system’s ability to effectively retrieve and process information from the outside of the system and from other system components, such as a memory (Wood et al., 2012).

Access to crucial information might be challenging in scenarios where an embodied AI system operates with restricted sensor capabilities or obstructed lines of sight. Consider a robot navigating a cluttered and dimly lit space. If its visual sensors are obstructed or limited, the system’s access to visual cues, such as identifying obstacles or determining the layout of the

environment, is compromised, and so is its awareness of the surroundings.

On the other hand, in a properly designed system, an embodied AI system can showcase effective access to information. For instance, in an autonomous vehicle equipped with advanced cameras, LiDAR, and radar systems, the system gains access to a rich set of data about its surroundings. At any time, the system has access to the data, and if one of these sensors fails, the system maintains awareness thanks to the redundancy of its senses.

#### 3.2 Information Integration

For a system to exhibit awareness of  $X$ , the integration of all data about  $X$  is essential. Integration goes beyond access as it involves putting together and synthesising information into a unified and meaningful representation.

Without access to the visual information, the robot in the example previously discussed finds itself lost in the environment. Having an alternative data source, such as a sonar, would alleviate the problem. However, the new system is susceptible to a new issue: the two sensors could provide contrasting data. If the robot fails to merge and integrate the information available into a coherent model of the environment, then it will not find itself in a better position than in the initial situation.

Similarly, the aforementioned robot has access to a diverse range of data sources about its surroundings. However, it needs to maintain a coherent model of the situation, to be aware of it. The integration of these diverse data sources empowers the AI system to navigate safely, showcasing a high level of awareness of its surroundings.

#### 3.3 Attention

Attention is a key component of awareness – or consciousness, depending on the discipline that is being considered. According to Taylor (Taylor, 2007), attention is the consciousness of a stimulus. It allows focusing on the most salient aspects while ignoring other *distractors*. Real-time processing is fundamental for maintaining awareness, and attention is one of the means to reduce the computational load of the system.

Trivially, any system with a sound design displays a certain form of *architectural* attention. Imagine a self-driving car: the system in charge of maintaining awareness of the surroundings will not receive data on which radio station is playing, by design. However, this is hardly a proper attention mechanism, as it boils

down to simply not having access to irrelevant data.

Instead, attention is about filtering out a part of the data, and the focus of attention can be limited to a handful of aspects at one time. Paying attention to a car several hundred meters behind while driving at high speed is not necessary, as the car's resources are better employed to detect obstacles in front of the vehicle. Nonetheless, the data about the car is still accessible, and the attention should be shifted towards it if, for instance, it turned on the police light bars signalling to make way.

### 3.4 Coherence

For embodied AI to demonstrate awareness of  $X$ , it must exhibit coherence in the decision-making associated with it. This involves maintaining consistency, both during the task at hand and over time.

To showcase awareness, the AI system must exhibit coherence with its own decision history. Memory, or by extension an internal model, is vital for this process. An embodied AI should maintain an accessible record of past decisions and outcomes, and produce consistent responses across similar scenarios, demonstrating the system's ability to apply past experiences to comparable situations. As a consequence, the system could be made able to learn from previous mistakes and predict the outcome of its actions.

In the same way, the system should be stable in its decisions during the execution of a task. It is expected that an autonomous car will suddenly reduce its speed when it detects an unforeseen obstacle. However, in a normal situation, the car is expected to maintain a constant speed showing awareness of the obstacles on the way.

### 3.5 Explainability

In the context of embodied AI awareness, explainability is a safeguard for safety and a means to accountability. The system must not only be aware of ( $X$ ) but also capable of elucidating its understanding and decisions regarding  $X$ . The explanation can be in any form, such as the English language or a diagram. It has however to be factual, representing the real motivations for a certain behaviour. Indeed, post-fact reasoning about the events that happened and why, which any Large Language Model certainly enables, does not add to the safety nor the accountability of the system.

It is easy to see how a factual explanation of why a self-driving car chose a specific course of action is essential for passengers, regulators, and other road users. For instance, in situations where the car over-

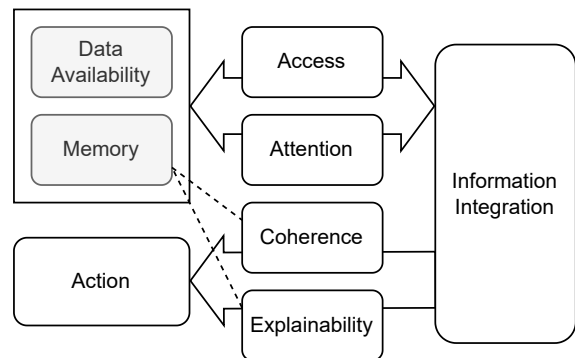


Figure 1: This diagram shows the relations between the awareness requirements presented.

rides human input or faces ambiguous road conditions, clear explanations ensure accountability and adherence to legal and ethical standards. While this aspect might appear as secondary, only a truly aware system can provide such an explanation.

### 3.6 Action

An integral aspect of embodied AI is the interaction with the external environment. Being aware of a certain aspect ( $X$ ) should be followed by a possible change of the internal behaviour or an intervention in the environment to bring about a desired change.

Consider an autonomous vehicle navigating a busy urban area. The system being aware of its surroundings is useless unless it can also modify its trajectory and speed to avoid obstacles, ensuring the safety of both passengers and others on the road. Intervening on the system's behaviour is not the only way to effect change: indeed, a system can also intervene in the external environment. For example, a smart building management system may adjust lighting and temperature based on occupancy patterns, to enhance energy efficiency and user comfort.

Without the ability for a behaviour change, reactive or proactive, the system's utility diminishes. Indeed, true awareness of a situation encompasses not just perceiving and understanding it but also adapting and responding effectively to its dynamics.

## 4 AWARENESS FOR EMBODIED AI

The requirements presented depend strongly on each other as shown in Figure 1. Access to internal and external information forms the foundational layer, allowing the system to perceive and collect data about its environment and have a memory. This information



is then subjected to information integration, where the system combines and synthesises data into a cohesive representation. Without access, the information integration process would have to rely passively on the data streaming from the environment. Attention acts as a dynamic filter, directing the system to focus on relevant aspects and optimising real-time processing.

Action completes the loop, as the system, thanks to the information integration, can dynamically interact with and influence its surroundings. Ensuring that the system's decisions and actions align with its understanding, and maintaining coherence over time is fundamental for successful and reliable interactions. Explainability serves as a critical component, demanding that the system accurately justifies its decisions, fostering transparency and accountability. Both these last functionalities require access, specifically to the memory of previous experiences.

Considering everything presented so far, we propose to call a system *aware* of *X* if:

- it has access to information about *X*, in the form of data availability, memory recall and forward modelling;
- it displays an attention mechanism towards *X*, filtering out distractors;
- it can successfully integrate available information into a model of *X*;
- it can act in response to *X*, changing its behaviour or intervening on the environment;
- it displays coherence in its decisions about *X*, with respect to its current and previous actions;
- it is explainable in its decisions about *X*, using verifiable data to justify them.

If the first measures are evidently necessary for a working system, it can be debated that action and explainability do not play a fundamental role. However, we argue that all the aspects presented are equally important for aware embodied AI.

In particular, the action is what distinguishes the aware system from a passive observer. Consider for instance a human without any motor capability: even without a possibility to change the state of the world that surrounds him, this subject is clearly still aware, as long as he can change his ideas and thoughts in response to external stimuli. However, if we know for certain that this is not the case, we would say that the subject is no longer aware.

On the other hand, a system that can interact with the external world – thus possessing the action requirement – but cannot explain the motivation behind its actions, cannot be defined as truly aware as it lacks the crucial element of transparency. Explainability

serves as the bridge between the system's internal processes and its external behaviour. Without the ability to articulate the reasons behind its actions, the system remains inscrutable, hindering our understanding and trust in its cognitive processes.

A relevant note is to be made, that in this work we borrowed the term *awareness* from studies revolving around humans, and we applied it to the world of machines. This was permitted by the similarities between the behaviour of an aware AI with the results of similar mechanisms taking place in humans and has nonetheless been done before (Drury et al., 2003; Holland, 2004; Schipper, 2014, just to cite a few). We want to underline that the scope of this definition is embodied AI, and it is not our intention to define awareness for humans and living creatures. We intentionally chose *awareness* to underline that what we want to achieve is a subset of *consciousness*, which remains a trait of mankind alone.

## 5 CONCLUSION

This paper initiates a conversation on practical strategies for enhancing awareness in embodied AI systems. We introduce six key elements as its foundations: access to information, information integration, attention, coherence, explainability, and action. Emphasising their interconnected and interdependent nature, we argue that these elements form a minimal base for constructing systems with heightened awareness.

While the proposed definition takes a practical approach to the topic, it's important to note that there is currently limited empirical evidence supporting it. The contribution underscores the need for future studies to validate and refine these insights, ensuring their effective implementation in real-world AI applications.

This contribution aims to spark a dialogue within the research community, fostering a dynamic exchange of ideas and perspectives. Our work aims not just to set a stage but to open a dialogue, inviting diverse voices to contribute to the evolving discipline of awareness in embodied AI.

## ACKNOWLEDGEMENTS

Funded by the Horizon Europe VALAWAI project (grant agreement number 101070930).

## REFERENCES

- Chella, A. and Pipitone, A. (2020). A cognitive architecture for inner speech. *Cognitive Systems Research*, 59:287–292.
- Chella, A., Pipitone, A., Morin, A., and Racy, F. (2020). Developing Self-Awareness in Robots via Inner Speech. *Frontiers in Robotics and AI*, 7.
- Chrisley, R. (2003). Embodied artificial intelligence. *Artificial intelligence*, 149(1):131–150.
- Dehaene, S., Lau, H., and Kouider, S. (2017). What is consciousness, and could machines have it? *Science*, 358(6362):486–492.
- Drury, J., Scholtz, J., and Yanco, H. (2003). Awareness in human-robot interactions. In *SMC'03 Conference Proceedings. 2003 IEEE International Conference on Systems, Man and Cybernetics. Conference Theme - System Security and Assurance (Cat. No.03CH37483)*, volume 1, pages 912–918 vol.1.
- Duan, J., Yu, S., Tan, H. L., Zhu, H., and Tan, C. (2022). A Survey of Embodied AI: From Simulators to Research Tasks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(2):230–244.
- Fodor, J. A. (1981). The mind-body problem. *Scientific american*, 244(1):114–123.
- Frith, C. and Frith, U. (2005). Theory of mind. *Current biology*, 15(17):R644–R645.
- Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., Martin, M., Nagarajan, T., Radosavovic, I., Ramakrishnan, S. K., Ryan, F., Sharma, J., Wray, M., Xu, M., Xu, E. Z., Zhao, C., Bansal, S., Batra, D., Cartillier, V., Crane, S., Do, T., Doulaty, M., Erapalli, A., Feichtenhofer, C., Fragomeni, A., Fu, Q., Gebreselasie, A., González, C., Hillis, J., Huang, X., Huang, Y., Jia, W., Khoo, W., Kolář, J., Kottur, S., Kumar, A., Landini, F., Li, C., Li, Y., Li, Z., Mangalam, K., Modhugu, R., Munro, J., Murrell, T., Nishiyasu, T., Price, W., Puentes, P. R., Ramazanov, M., Sari, L., Somasundaram, K., Southerland, A., Sugano, Y., Tao, R., Vo, M., Wang, Y., Wu, X., Yagi, T., Zhao, Z., Zhu, Y., Arbeláez, P., Crandall, D., Damen, D., Farinella, G. M., Fuegen, C., Ghanem, B., Ithapu, V. K., Jawahar, C. V., Joo, H., Kitani, K., Li, H., Newcombe, R., Oliva, A., Park, H. S., Rehg, J. M., Sato, Y., Shi, J., Shou, M. Z., Torralba, A., Torresani, L., Yan, M., and Malik, J. (2022). Ego4D: Around the World in 3,000 Hours of Egocentric Video. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18973–18990.
- Holland, O. (2004). The Future of Embodied Artificial Intelligence: Machine Consciousness? In Iida, F., Pfeifer, R., Steels, L., and Kuniyoshi, Y., editors, *Embodied Artificial Intelligence: International Seminar, Dagstuhl Castle, Germany, July 7-11, 2003. Revised Papers*, Lecture Notes in Computer Science, pages 37–53. Springer, Berlin, Heidelberg.
- James, W. (1890). The principles of psychology, vol. 1. Henry Holt and Co. New York.
- Pfeifer, R. and Bongard, J. (2006). *How the body shapes the way we think: a new view of intelligence*. MIT press.
- Schipper, B. C. (2014). Awareness. Available at SSRN 2401352.
- Taylor, J. G. (2007). Through machine attention to machine consciousness. In Chella, A. and Manzotti, R., editors, *Artificial Consciousness*, pages 24–47. Imprint Academic.
- Wood, R., Baxter, P., and Belpaeme, T. (2012). A review of long-term memory in natural and synthetic systems. *Adaptive Behavior*, 20(2):81–103.