EMERGE

WP1 Conceptual framework

D1.1 Local awareness criteria

Version: 1.0

Date: 31/03/2023

Funded by the European Union

European Innovation Council

Funded by the European Union under Grant Agreement 101070918. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Innovation Council and SMEs Executive Agency (EISMEA). Neither the European Union nor the granting authority can be held responsible for them.

Document control

Project title	Emergent awareness from minimal collectives
Project acronym	EMERGE
Call identifier	HORIZON-EIC-2021-PATHFINDERCHALLENGES-01-01
Grant agreement	101070918
Starting date	01/10/2022
Duration	48 months
Project URL	http://eic-emerge.eu
Work Package	WP1 Conceptual framework
Deliverable	D1.1 Local awareness criteria
Contractual Delivery Date	M6
Actual Delivery Date	M6
Nature[1]	R
Dissemination level[2]	PU
Lead Beneficiary	LMU
Editor(s)	Ophelia Deroy (LMU)
Contributor(s)	Nadine Meertens (LMU), Bahador Bahrami (LMU)
Reviewer(s)	Davide Bacciu (UNIPI)
Document description	This report identifies 6 challenges for the project of implementing awareness in artificial agents - both at the conceptual, engineering, ethical and industrial levels - and offers a dimensional approach as a way to address them.

[1]R: Document, report (excluding the periodic and final reports); DEM: Demonstrator, pilot, prototype, plan designs; DEC: Websites, patents filing, press & media actions, videos, etc.; DATA: Data sets, microdata, etc.; DMP: Data management plan; ETHICS: Deliverables related to ethics issues.; SECURITY: Deliverables related to security issues; OTHER: Software, technical diagram, algorithms, models, etc.

[2]PU – Public, fully open, e.g. web (Deliverables flagged as public will be automatically published in CORDIS project's page); SEN – Sensitive, limited under the conditions of the Grant Agreement; Classified R-UE/EU-R – EU RESTRICTED under the Commission Decision No2015/444; Classified C-UE/EU-C – EU CONFIDENTIAL under the Commission Decision No2015/444; Classified S-UE/EU-S – EU SECRET under the Commission Decision No2015/444

Funded by the European Union
European Innovation Council
Funded by the European Union under Grant Agreement 101070918. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Innovation Council and SMEs Executive Agency (EISMEA). Neither the European Union nor the granting authority can be held responsible for them.

Version control

Version[3]	Editor(s) Contributor(s) Reviewer(s)	Date	Description
0.1	Ophelia Deroy	18.03.2023	TOC proposed by editor
0.2	Davide Bacciu	23.03.2023	TOC approved by reviewer
0.4	Ophelia Deroy	25.03.2023	Intermediate document proposed by editor
0.5	Davide Bacciu	26.03.2023	Intermediate document approved by reviewer
0.8	Ophelia Deroy	27.03.2023	Document finished by editor
1.0	Davide Bacciu	31.03.2023	Document released by Project Coordinator

[3] 0.1 – TOC proposed by editor; 0.2 – TOC approved by reviewer; 0.4 – Intermediate document proposed by editor; 0.5 – Intermediate document approved by reviewer; 0.8 – Document finished by editor; 0.85 – Document reviewed by reviewer; 0.9 – Document revised by editor; 0.98 – Document approved by reviewer; 1.0 – Document released by Project Coordinator.

Funded by the European Union
European Innovation Council

Funded by the European Union under Grant Agreement 101070918. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Innovation Council and SMEs Executive Agency (EISMEA). Neither the European Union nor the granting authority can be held responsible for them.

Abstract

The term “awareness” has no single definition in biological or artificial agents. Anticipating that the issue magnifies when considering collectives of artificial agents, we here provide a systematic overview of the problems that already arise at the level of single agents. The core challenge identified in EMERGE is that existing frameworks are ineffective or vague in explaining, facilitating, and supporting cooperative behaviours in artificial agents. The lack of a compelling theory of global awareness in AI is currently a significant barrier to the effective deployment of artificial agents in the real world.

The first step in tackling this grand challenge is to distinguish six challenges that a satisfactory concept of awareness should address (justification, explanatory role, meaning, metrics, implementation and ethics). EMERGE proposes a novel dimensional account inspired by the dimensional accounts of consciousness offered to replace the unidimensional “level” metaphor dominating the clinical neuroscience literature. The dimensional account presented here is less committed to the exact labels of dimensions than it is to account for differences in contents (i.e. what the agent is aware of) as well as for the interdependence of specific contents usually captured in the level metaphor (e.g. to be self-aware of one’s location depends on being aware of space).

Funded by the European Union
European Innovation Council

Funded by the European Union under Grant Agreement 101070918. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Innovation Council and SMEs Executive Agency (EISMEA). Neither the European Union nor the granting authority can be held responsible for them.

Disclaimer

This document does not represent the opinion of the European Union or European Innovation Council and SMEs Executive Agency (EISMEA), and neither the European Union nor the granting authority can be held responsible for any use that might be made of its content.

Neither the EMERGE consortium as a whole, nor a certain party of the EMERGE consortium warrant that the information contained in this document is capable of use, nor that use of the information is free from risk and does not accept any liability for loss or damage suffered by any person using this information.

Acknowledgement

This document is a deliverable of EMERGE project. This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement Nº 101070918.

Table of contents

Contents

Funded by the European Union
European Innovation Council
Funded by the European Union under Grant Agreement 101070918. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Innovation Council and SMEs Executive Agency (EISMEA). Neither the European Union nor the granting authority can be held responsible for them.

List of tables

List of figures

Funded by the European Union
European Innovation Council
Funded by the European Union under Grant Agreement 101070918. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Innovation Council and SMEs Executive Agency (EISMEA). Neither the European Union nor the granting authority can be held responsible for them.

Introduction

Artificial agents and robots are designed to interact with their physical environment, which is often subject to unpredictable changes and factors beyond the agent's control. As the number of devices and systems increases, central control becomes less feasible. In rapidly evolving situations, waiting for a response from a central controller may not be an option, so it's essential to ensure that artificial agents can operate more locally and with more autonomy (even though they may not have complete autonomy and keep humans in the loop). Awareness has been proposed as a way to enhance their efficiency, resilience, and flexibility, allowing them to adapt to unforeseen situations and operate continuously.

If bringing awareness to AI appears to be a goal or a solution, what does this solution demand? Here we map the issues that computer scientists, other disciplines, and users can encounter along the way when looking to turn awareness into a technological reality. In this report, we distinguish six different problems - each of which raises different challenges for different communities - computer scientists, engineers, users, cognitive scientists, philosophers and ethicists who have vested interest and expertise in awareness.

The first problem we can identify is a problem of justification: why is using the term "awareness" a good idea, and wouldn't other words be as appropriate? The second is the explanatory role: what is the value of explaining the performance or functioning of an artificial agent (or set of agents) by referring to awareness? The third issue concerns the multiple uses of the term "awareness": what are the various meanings that the concept covers, and how do they relate to each other? The fourth and fifth problems concern ways to measure and implement awareness, while the sixth concerns the ethical implications of having aware systems.

Table 1: Awareness is offered as a solution to flexibility and other problems, but raises its own problems. This table shows six different problems with their accompanying challenges and the solutions that we propose.

Problem	Challenge	Main community of concern	Step to solution provided in this paper
Justification	Getting rid of the term; using other terms	Philosophers, users	Dimensional concept of artificial awareness
Explanatory role	No benefit to users or explanation of the system's behaviour	Computer scientists, users	Distinction between explanation to user and functional role
Meaning	Too many disparate meanings; loose talk	Computer scientists, cognitive scientists, ethicists	Taxonomy; CML dimensional model
Metrics	No way to assess and measure	Computer scientists, engineers, users	Specific metrics vs. Awareness as modulating other metrics
Implementation	Speculative; no plausible computational or hardware solution	Computer scientists, engineers	
Ethics	Responsibility gap	Users, ethicists, companies	

Funded by the European Union
European Innovation Council
Funded by the European Union under Grant Agreement 101070918. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Innovation Council and SMEs Executive Agency (EISMEA). Neither the European Union nor the granting authority can be held responsible for them.

1. Justifying the concept of awareness

1.1. Awareness vs. Consciousness

Alan Turing and John von Neumann, the founders of the modern science of computation, entertained the possibility that machines would ultimately mimic all of the brain's abilities, including consciousness. The idea that machines or robots could develop consciousness continues to be a fringe view, which looks somewhat like a vestige of Pygmalion's fantasy or speculative science fiction (e.g. Signorelli, 2018). The word "consciousness", like many pre-scientific terms, is used in widely different senses and has led to widely different theoretical understandings across philosophy and neuroscience.

The concept of awareness does not appear as demanding or mysterious as consciousness. Though it is sometimes used as one of its synonyms, one of its advantages is to be more detached from everyday associations with qualitative consciousness - what is philosophically referred to as "something-it-is-like-to-be-x", following Nagel (1974). Relatedly, while having a qualitative experience seems to be an all-or-nothing (either one has it or not, and one cannot have some degree of qualitative experience), being aware is more welcoming of grades (one can be more or less aware of something). While we take this later feature to be a critical reason to stick to the concept of awareness, one inconvenience is that the concept still lacks a proper definition both in biological and artificial systems.

The lack of a clear definition is acknowledged in the computer science community but leads to different responses. A pragmatic stance, illustrated for instance by David Levy (Levy, 2009, p. 210), claims that it is sufficient to have a general agreement about what we mean by awareness and suggests "let us simply use the word and get on with it.".

Others choose to settle for a prototypical profile of what is being aware without providing a set of necessary or sufficient features (Figure 1b). For instance, Chatila et al. (2018, p. 1) consider relevant: "… the underlying principles and methods that would enable robots to understand their environment, to be cognisant of what they do, to take appropriate and timely initiatives, to learn from their own experience and to show that they know that they have learned and how." In this prototype approach, a system would be called more or less aware depending on loose family resemblance with this prototype.

Both the pragmatic and prototypical approaches fall short of providing a clear answer to how awareness should be measured and therefore compared across systems. Both may explain the remaining scepticism that surrounds the use of "awareness" in engineering and computer science. Turning to philosophy and neuroscience, we find it is possible to find other approaches that are more rigorous and allow for a comparison between systems.

Another approach, indeed dominant in cognitive neuroscience and philosophy of consciousness, consists in distinguishing different kinds of awareness. Two examples borrowed here from the literature on consciousness can illustrate this approach: writing about machine consciousness, Dehaene et al. (2021), for instance, distinguish global availability from self-monitoring, which is the capacity to have self-referential thought and report on one's states. Another example is Ned Block's distinction between phenomenal consciousness and access consciousness: Whereas phenomenal consciousness relates to the experience, to what it is like to be in a conscious mental state, access consciousness refers to a mental state's availability for use by the organism, for example in reasoning and guiding behaviour, and describes how a mental state is related with other mental states.

Funded by the European Union
European Innovation Council
Funded by the European Union under Grant Agreement 101070918. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Innovation Council and SMEs Executive Agency (EISMEA). Neither the European Union nor the granting authority can be held responsible for them.

These approaches make it, at least in principle, possible for one kind of consciousness to occur without the other: in the two accounts mentioned above, a state can be globally available but not self-reported (e.g. blindsight) or a state can be phenomenally conscious without being accessed (Block, 2011). This said, we can distinguish between two versions: If the two kinds or features are mutually exclusive, the approach provides a taxonomy or classification, for instance, between two kinds of awareness that are independent (Figure 1c). If one kind of awareness is more fundamental in that it can occur without the other, but having the two kinds together is a fuller realisation, this leads to a "building block" or pyramidal approach (Figure 1d).

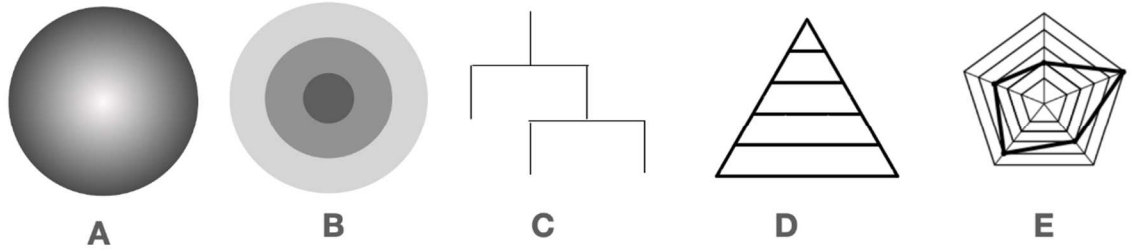

Figure 1: Five approaches to awareness in AI. (A) pragmatic approaches take the semi-scientific concept as it is, without scrutiny (B) prototype approaches consider a core example of what an aware system is or should be and map family resemblance between this core example and other instances (C) taxonomies distinguish kinds or features of awareness, which are independent of each other (D) pyramidal approaches distinguish kinds or features of awareness but accept that some kinds (or features) depend on the existence of some other more basic kinds or features (E) dimensional approaches do not commit to distinct between kinds or features, but consider that different aspects can exist at different degrees in either fully dependent or more dependent manners. Dimensional approaches commit to each dimension being graded.

A more novel approach, which we are developing as part of EMERGE, provides a dimensional account of awareness. Like the previous two accounts and against the pragmatic account, the approach provides clarity on the concept. Where it differs from the prototypical approach is to stress the need to organise similarities and differences between aware systems in a systematic and quantitative manner. If the goal of developing artificial awareness comes hand in hand with the need to measure awareness and be able to provide comparisons between systems, we suggest that having different dimensions and ways to assess each one separately is a more fruitful approach (see 4. Below)

The difference between the “building block” or hierarchical approach, but also with related proposals in cognitive neuroscience (Bayne et al., 2016) is not only to be more fine-grained but to be able to capture both the independence or interdependence between different aspects of awareness: this can be done using the difference between integral and separable dimensions found in psychophysics. The distinction, when used phenomenologically, is “...between dimensions which can be pulled apart, seen as unrelated, or analysable, and those which cannot be analysed but somehow are perceived as single dimensions” (Garner & Felfoldy, 1970, p. 225). An example of a pair of separable dimensions is colour and shape; an example of a pair of integral dimensions is Munsell chroma and value (i.e., saturation and lightness). Additionally, the dimensional framework enables the distinction between integral, co-dependent features, and configurational dimensions, which depends on certain patterns of co-occurrence between features (e.g. symmetry) (Fig. 2).

Funded by the European Union
European Innovation Council
Funded by the European Union under Grant Agreement 101070918. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Innovation Council and SMEs Executive Agency (EISMEA). Neither the European Union nor the granting authority can be held responsible for them.

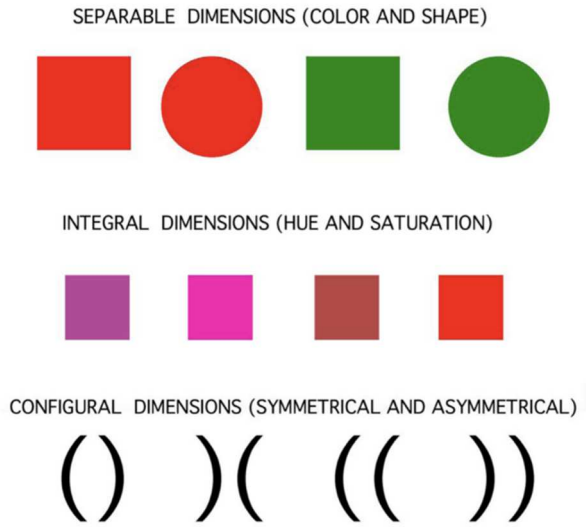

Figure 2: Three types of dimensions. Separable dimensions are fully independent; integral dimensions are dependent or co-instantiated; configured dimensions emerge on several features. The distinction can be transposed to dimensional models of awareness.

1.2. Humans tend to attribute awareness to AI systems

As described above, there are multiple ways in which the concept of awareness can be analysed so as to become valuable and measurable in computer science and engineering. While this is a scientific reason to select the concept, there are other reasons, this time coming from the naive psychological attitudes of users. There is abundant evidence that humans attribute human or animate features to AI - including awareness. While anthropomorphism is a genuine bias, there is evidence that attributions to AI differ significantly from attributions to animals or humans (though these conclusions still ask for more research and support). The question remains empirically complex, not just because of the variety of AI systems that exist but also because the cognitive processes by which humans attribute mental characteristics to others (encompassed as social cognition) are also complex.

The extreme case of social robots is the one most often studied by psychologists and experts in human-AI interactions. Social robots have several characteristics that make them special for humans: Besides being capable of limited decision-making and learning, they can exhibit behaviour and interact with people. In addition, capabilities like nonverbal immediacy of robot social behaviour (Kennedy et al., 2017), speech recognition and verbal communication (Grigore et al., 2016), facial expression, and the perceived "personality" of robots play important roles in how humans respond to robots.

Consequently, humans tend to develop unidirectional emotional bonds with robots, project lifelike qualities, attribute human characteristics, and ascribe intentions to social robots (e.g. Gunkel, 2018).

Regarding human-looking robots, behavioural studies suggest that people interact with them as if they were human agents rather than mere machines - at least if they look sufficiently human (for an overview, see Broadbent, 2017). For instance, people apply stereotypical social categories such as gender or (out)group memberships (e.g. Złotowski et al., 2015; Salles et al., 2020) to robots. They also display typical social behaviour toward them: they punish a robot that admits wrongdoing (Lee et al., 2021) and accept its apologies. Further on the moral side, evidence shows we tend to recognise AI as partly accountable for their mistakes (Kahn

Funded by the European Union
European Innovation Council
Funded by the European Union under Grant Agreement 101070918. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Innovation Council and SMEs Executive Agency (EISMEA). Neither the European Union nor the granting authority can be held responsible for them.

et al., 2012). Various studies more specifically suggest that people are ready to use naive psychological categories - including categories supposing awareness such as "knowing", "believing", "being responsible", etc. -when explaining the behaviour of robots. For instance, Graaf and Malle (2019) provided people with verbal descriptions of robot and human behaviours across different contexts and asked them to explain why the agent had performed them. People used the same conceptual toolbox of behavioural explanations for human and robot agents.

However, an equally substantial number of studies using interactive set-ups suggest that naive users draw a difference between AI and humans: they notably trust AI less in some conditions (Burton et al., 2017), will reciprocate less towards an AI than a human stranger (Karpus et al., 2021) and show less empathy and reciprocity (Mamoodi et al., 2018). Various labels have been used to characterise these tendencies, including "algorithm aversion" and "algorithm exploitation". The question is whether this has to do with attributions of awareness.

Following Gray, Gray, and Wegner (2007), we suggest that the attributions of human characteristics to AI should not confuse the attribution of a capacity for agency and the attribution of a capacity for experience (such as hunger, fear, pain, etc.). For instance, lay people attribute a high degree of experience to a baby but no agency; they also attribute high agency but no experience to a deity. To determine if robots are considered human agents, assessing the degree to which agency and experience are ascribed to them is important. In a survey with 184 students, however, the answers to the question "Do you believe that contemporary electronic computers are conscious?" were: No: 82%; Uncertain: 15%; Yes: 3% (Reggia et al., 2015). It is worth highlighting, however, that the question in the survey was about "contemporary electronic computers" and not AI or robots.

While evidence that AI is granted the explicit ability to plan and act and elicit the same action representation mechanisms as human interaction partners is strong (e.g. Chaminade et al., 2010), the same is not true for the ability to experience or have other mental states. For instance, interpreting a human's gaze or predicting their actions takes less time than doing so for humanoid robots. This advantage for processing the human gaze was observed in tasks that require representing others' minds rather than tasks that focus on detecting a change in others' gaze. This suggests that people can extract non-mentalistic information from a robot's gaze, but they are less able to infer the future actions of robots than humans.

Neuro-imagery studies confirm that mechanisms linked to the attribution of sentience or mental states are not fully activated or are less activated when interacting with humanoid robots. For example, increased neural activity was observed in mentalising areas (such as the temporoparietal junction and dorsal prefrontal cortex) during human-human interactions but not during human-robot interactions This was observed during eye contact but also conversation (Rauchbauer et al., 2019). Even when AI is provided "a human face", viewing robotic facial expressions evokes less activity in mentalising areas than viewing human facial expressions (Hmamouche et al., 2020).

2. The role of artificial awareness

What advantage does it bring to speak of a single robot or eventually of a collective of robots (i.e. swarm) as being aware of a change in the surrounding environment or context, rather than saying that it has information about the changes of context or can respond to changes of context? This is a second core challenge, as it asks how the attribution of awareness to the

Funded by the European Union
European Innovation Council
Funded by the European Union under Grant Agreement 101070918. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Innovation Council and SMEs Executive Agency (EISMEA). Neither the European Union nor the granting authority can be held responsible for them.

robot differs and is better than the attribution of mere information or capacity to respond (input-output).

2.1. Awareness by design

The first point to stress is the difference between the problems of identifying and accounting for awareness in animals (which are attributions problems) and the problems of positing or designing awareness in AI (which are design-led). Most sound accounts attributing awareness in non-human animals tend to be done because awareness is supposed to be necessary or at least useful to explain certain existing behaviours. For instance, different forms of agentive awareness can be posited to explain why certain animal species exhibit more or less behavioural flexibility, communication or control. The problem with taking this observer's stance with AI is that it plays on what we call Pygmalion's fallacy: the idea that awareness could be an ingredient inside the machine that the designer had no initial knowledge of or control over and is trying to infer afterwards. While this can work for naturally evolved biological agents, where awareness could have evolved without us being consulted or contributing, the same is not true for AI.

So the first difference is that artificial awareness is linked to promoting a specific function implemented by design. To make the contrast more salient, consider the difference between reptiles and moving robots. We look at a lizard and test whether it is sensitive to the context in a detour task. Based on the lizard's performance, we then infer whether it is aware of various contextual features of its environment that help it attain its goals. For the robot, we want to make sure that it is sensitive to various contextual features of its environment in order to allow or let it learn to reach a certain level of performance. In turn, its performance indicates the robot's awareness of the contextual features of the environment in which it operates to attain its goal.

This distinction means that the question of awareness in artificial systems can fit into two different kinds of explanatory roles. In the case of animals, the explanatory goal is to determine whether the animal is really aware. Biological sciences, like physical sciences, operate primarily in a realist stance - where the entities posited by the theory are supposed to correspond to the realities that exist in the natural world (this said, some anti-realist stances are also defended and accept that the entities posited by theories are simply fictions that are useful to make sense or predict the natural world). The difference between realist and anti-realist stances can be mirrored in the artificial world by a difference between a descriptive stance and a mere explanatory stance. According to the descriptive stance, awareness is really present (at different degrees or in different ways) in the system and plays a functional role. If awareness (or one of its dimensions in our framework) goes missing, we expect different behaviour. For instance, in comatose patients and disorders of awareness, certain behaviours - like the capacity to report or self-monitor - are degraded or missing. According to an explanatory stance exhibited in the X-AI literature, awareness is simply a different way to speak of complex input-output functions, which we have no other or better ways of describing. The explanatory stance is closer to a proper understanding of what a "Turing test" approach to awareness would be: if we cannot distinguish, for instance, between a biological agent which we have independent reasons and evidence to consider aware (or aware to a certain extent) and an artificial agent, then we have sufficient reasons to talk as if the artificial agent is aware, but we are not committed to more. In the Turing test, for instance, the fact that it becomes impossible for a human to distinguish between a machine and a human does not make the machine human. It makes it not even human-like (in essence) but superficially similar (for people like us) to a human. The difference between the two stances seems theoretical but

Funded by the European Union
European Innovation Council
Funded by the European Union under Grant Agreement 101070918. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Innovation Council and SMEs Executive Agency (EISMEA). Neither the European Union nor the granting authority can be held responsible for them.

has consequences for ways of thinking about the measurement or ethics of awareness of AI (as discussed below).

2.2. What awareness explains

Below we provide a non-exhaustive list of behaviours that are taken to require the presence of awareness or can be better explained by positing awareness. Some of these behaviours or manifestations remain based on human equivalents. The classification is also based on the broad difference between the availability/access function of awareness vs the self-monitoring functions of awareness.

Access and agent-level availability

- The same input leads to different "inner" states of the system.
- Logical dependence on inner states rather than the input (Attentional blink: conscious perception of item A prevents the simultaneous perception of item B; planning dependent on inner states)
- All-or-none selection and broadcasting of relevant content (e.g. Conscious perception of a single picture during visual rivalry; Conscious perception of a single detail in a picture or stream)
- Flexible routing of information
- Sequential performance of several tasks
- Flexible goal-directed behaviour
- Model-based learning
- Ability to make motivational trade-offs (Ginsburg & Jablonka, 2019)
- Integrative, multimodal representations
- Susceptibility to illusions

Self-monitoring

- Self-measurements
- Self-quality control
- Self-monitoring (PID adaptation)
- Self-diagnosis (status update)
- Self-analysis (self-healing)
- Self-learning
- Self-prognosis (self-planning)
- Self-optimisation (self-configuration)
- Self-location and orientation
- Self-belonging (assembly)

3. The Contents-Modes-Locus (CML) framework: Working towards dimensions of awareness.

3.1. State of the art: contents, levels and locus.

Coming back to the types of approaches listed in Figure 1, certain approaches (pragmatic, taxonomy) to awareness in artificial agents seem to recommend categorical answers to the question of whether a system is aware: the answer that one expects is whether awareness is present or not (and which type of awareness). These approaches also tend to provide

Funded by the European Union
European Innovation Council
Funded by the European Union under Grant Agreement 101070918. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Innovation Council and SMEs Executive Agency (EISMEA). Neither the European Union nor the granting authority can be held responsible for them.

responses to "what is the system aware of?" rather than to "how aware is the system?" for a given content or across all sorts of content.

The cognitive neuroscience and philosophy of awareness here provide a very different angle and focus at least as much on contents as they do on levels of awareness - notably because of a concern for patients and altered states seen during drug experiences or anaesthesia. The approach in terms of 'levels' of awareness or consciousness echoes the pyramidal models discussed in part 1. The idea of levels is more clearly conceptualised in the field of consciousness research, where it is used to refer to states of consciousness that are linked to certain clinical conditions, such as post-coma disorders. The same idea of levels of consciousness has also been extended to include sedation, sleep, and epileptic absence seizures, as well as human infancy and non-human animals, where lower levels of consciousness are suggested to occur.

A unified taxonomy for states of consciousness has been proposed, which suggests that these states can be measured along a single dimension of scalability, more or less understood at a difference in the overall amount of awareness that introduces some categorical distinctions in what the patient is capable of doing.

The classification suggests that patients in minimally conscious states [MCS] have a higher level of consciousness than those in a vegetative state and that Emerging from Minimal Conscious States patients have a higher level of consciousness than MCS patients [e.g. Bruno et al., 2011). However, the single-dimensional approach to understanding consciousness does not capture the full complexity of global states of consciousness (Bayne et al., 2016) and leaves, for instance, many comparisons problematic, be it across humans (How does the level of consciousness observed in sleep differs or resembles the level of consciousness in minimally conscious states?) or across species (How does the level of consciousness present in a dog compare to the level of consciousness observed in a sedated patient?)

A different question which may be distinctive of artificial agents, has to do with "where" awareness occurs. There, the terminology used by some researchers uses the word "levels" (central or global level vs Distributed and local level). Still, it is better to keep the idea of locus because one system cannot be in two levels of awareness (a patient cannot be both in a vegetative state and a minimally conscious state). In contrast, certain systems could have two loci of awareness for different contents (local awareness of space vs global awareness of goal, for instance). Central awareness generally occurs outside the agent - in a control system. Distributed or decentralised situational awareness allows a group of robots to efficiently gather information about their surroundings and respond accordingly without needing a central hub for data storage or control. Each robot generates its own local awareness, which informs its actions in real-time. This approach, which we pursue in EMERGE, allows robots to utilise low-cost sensors and communication tools with limited range, like cameras, distance sensors, and Bluetooth. As a result, a collective situational awareness can emerge from combining each robot's local data, enabling the swarm to act as a cohesive unit across various applications, from environmental monitoring to logistics (e.g. Jones et al., 2020). By relying on local data, communication and processing requirements are reduced. Although decentralised systems are often seen as challenging to control, they enable users to interact with individual robots and the swarm as a whole, making it easier for non-experts to install and operate the robots.

The literature on awareness in smart systems sometimes uses the level idea, but the concepts used in the list of levels tend to combine differences in content, locus, and what we would as functions. A good example of such a mix is Duffy (2016), who defines the following awareness levels.

Funded by the European Union
European Innovation Council
Funded by the European Union under Grant Agreement 101070918. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Innovation Council and SMEs Executive Agency (EISMEA). Neither the European Union nor the granting authority can be held responsible for them.

- Level 1: Adaptive - Automatically adapts to changes in the environment like regular PID controller.
- Level 2: Property Aware - a semantic interpretation and attribution of monitored data based on expectations and goals. A system that monitors its own properties is self-aware.
- Level 3: History Aware - The system maintains a history of observations. Changes over time are monitored and assessed.
- Level 4: Predictive - A system with the capability to simulate if-then scenarios is called predictive
- Level 5: Group aware - Besides the self and the environment, the system recognises a peer group with shared goals and/or similarities in behaviour.

Another example is situation awareness, as formulated by Endsley. The classical definition of situation awareness describes it as: "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning and the projection of their status in the near future" (Endsley, 1988, p. 97). Endsley formulated this as such with the aviation domain in mind. It describes how all elements of the aviation team need to come and work together to achieve situation awareness (Endsley, 2013). While on the face of it, situational awareness considers different contents (being aware of the environment, of meanings and of future plans), the concept also establishes a hierarchy as being aware of the environment and meanings is required to have an awareness of future plans (also called predictive awareness).

Our recommended dimensional model acknowledges that there are different domains where awareness needs to be defined and measured - i.e. contents, levels and locus. It captures the differences by integrating content-related dimensions (what the system is aware of) and their possible interdependence (for instance, self-awareness may require awareness of space and time, and therefore represent a higher level of awareness), and considering how much awareness of these contents exists at each relevant locus of awareness (centralised, distributed).

3.2. Matching contents of awareness into separable and interdependent dimensions

The literature on awareness generally focuses on the contents of awareness, or in other terms: what one might be aware of. A brief overview of the existing literature allows us to make a preliminary distinction between the various contents of awareness that are distinguished, but also their interdependence.

3.2.1. Goal and task awareness

A central part of artificial intelligence and smart robots lies in accomplishing a task or goal. The difference between having a goal and being aware of a goal (sometimes labelled as task awareness) can easily become terminological, and to avoid this, one should make sure that awareness comes with specific conditions. One of them comes from flexible adjustment to optimise the task in changing conditions. Another one can come from the transfer of learning from one task to another. For instance, the rapid acquisition of new skills in humans is facilitated by the utilisation of prior knowledge - which itself is supposed to rely on the awareness of tasks and goals. When learning a skill that is connected to a broad range of previously mastered skills, relevant knowledge from prior skills can be recalled and applied to expedite the acquisition of the new skill. To illustrate, suppose we are learning a new snowboarding trick, drawing from our fundamental snowboarding knowledge, skiing expertise,

Funded by the European Union
European Innovation Council
Funded by the European Union under Grant Agreement 101070918. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Innovation Council and SMEs Executive Agency (EISMEA). Neither the European Union nor the granting authority can be held responsible for them.

and skateboarding experience. By exploiting our fundamental snowboarding knowledge and integrating inspiration from our skiing and skateboarding experience, we can rapidly master this feat. Recent progress in meta-learning in AI has aimed to achieve a similar thing and provide machines with a way to swiftly adapt to a new task using only a few examples. The approach involves first training an internal representation that matches similar tasks. Such representations can be learned by examining a distribution over similar tasks as the training data distribution. Model-based meta-learning methods propose to identify the task identity from a few sample data, regulate the model's state (e.g., RNN's internal state or external memory) using the task identity, and make appropriate predictions with the adjusted model (Vuorio et al., 2019).

3.2.2. External awareness

A central term that comes up in the artificial intelligence literature is context awareness. This is generally used to refer to applications that react to the information they sense from the environment - relating more to environmental awareness mentioned above with Endsley - rather than just operating on data provided to them (Benerecetti et al., 2001). Information here can be things such as location, time, objects, and agents in the environment.

This said, meaning and categorisation according to goals seem also to be part of context awareness. Intuitively, being aware of objects in the environment comes down to the capacity to assign properties to them, with the same thing applying to events. Spatial information on its own is insufficient for this (Chou et al., 2012). What this shows still is the assumption that environmental awareness is foundational and that meaning awareness is at another level.

The descriptions of awareness, as formulated above, could perhaps be further unpacked by looking at several of the elements they describe a bit more closely. Two elements that come up in almost all these descriptions and would be relevant for all of them are the awareness of space and time.

Spatial awareness has no unique definition, and there is no consensus on what it might mean for a being to be aware of space or, perhaps more aptly, be aware within space. The question of whether we need an awareness of space in absolute terms in order to perceive objects or properties as spatial is not settled (e.g. Schwenkler, 2012). Moreover, there seems to be a link present between spatial awareness and object perception (Campbell, 2017).

The literature on what we might call temporal awareness is also limited. However, one clear distinction can be made at the outset based on the philosophical literature on time, namely between physical spacetime and psychological time (Brown, 1990). In broad strokes, one might term these as the way in which time (static) actually is and the manner in which we perceive it from a subjective human point of view (dynamic) (Dainton, 2013). In AI, time awareness is an open discussion in that there might be a need for a standard time to be established within a distributed network in order to allow the network and individuals within it to know when an event occurred and compare data effectively (Hwang, 2019). But beyond this mere sharing of a standard time, it might also be relevant for AI systems to have information more akin to our psychological view of time (Maniadakis et al., 2009). This is where elements such as continuity, duration, simultaneity, persistence, change, succession, and an experience of past, present and future come in.

3.2.3. Internal and self-awareness

Introspective awareness relates to the awareness of inner states: Internal states can be made up of emotion, belief, desire, intention and expectation, or they can be processes such as sensation, perception, conception, simulation, action, planning and thought. An important

Funded by the European Union
European Innovation Council
Funded by the European Union under Grant Agreement 101070918. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Innovation Council and SMEs Executive Agency (EISMEA). Neither the European Union nor the granting authority can be held responsible for them.

question is whether goal awareness is already an instance, albeit a minimal form, of introspective awareness.

Metacognitive awareness sometimes includes general forms of introspective awareness (being aware of the state one is in) but tends to add an evaluative dimension to this introspective awareness in the form of an assessment of one's probability of being correct regarding a given task (being subjectively confident that one's inner state is correct). Metacognitive awareness requires introspective awareness and therefore adds not just a distinct content but also a distinct level.

Another internal content of awareness that is distinguished in the literature about human awareness is self-awareness, which usually differs from introspective awareness because it focuses on being aware of the continuity of one's body or unity through time.

Here a distinction is made between bodily self-awareness and reflective self-awareness. Bodily awareness entails being aware of your body as something distinct from the environment and an ability to distinguish internal from external input (Zhao, 2018; Bermúdez, 2011). For instance, a robot can continuously create a concept of its own physical structure (body self-modelling) and uses this self-model to generate forward locomotion with four legs initially without knowing what its body actually looks like. When the robot's structure changes unexpectedly, it can reform its internal self-model to generate new behaviours to compensate for and accommodate these changes (Bongard et al., 2006). In this case, it remodels the concept of its own physical structure to generate forward locomotion with three legs when one of its legs is removed.

Another content of self-awareness mentioned in the literature on artificial intelligence is agentive self-awareness, which is related to the capacity for having first-person representations of bodily/physical actions undertaken (Farina, 2022). This content usually presupposes bodily self-awareness.

Reflective self-awareness generally describes the ability to see yourself in the light of others or, in more general terms seeing yourself as a subject (Zhao, 2018). Takiguchi et al. (2013) explain how self-awareness is needed for autonomous robots in order for them to be capable of placing themselves in a better position to achieve their tasks. Related to these contents would be proprioception (Gallagher, 2007) and interoception.

3.2.4. Interactive awareness

The literature on human-ai interactions has also focused on other contents of awareness. A few to think of would be workspace awareness (who is working on what in the shared workspace) and the equivalent of social awareness (understanding connections within the group (see, e.g. Drury et al., 2003). Some systems may be aware of both individual tasks and shared tasks, or only of one - which means that these contents fit on different dimensions. This form of awareness of group-level events and other agents should not be mistaken for the collective locus of awareness.

Funded by the European Union
European Innovation Council
Funded by the European Union under Grant Agreement 101070918. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Innovation Council and SMEs Executive Agency (EISMEA). Neither the European Union nor the granting authority can be held responsible for them.

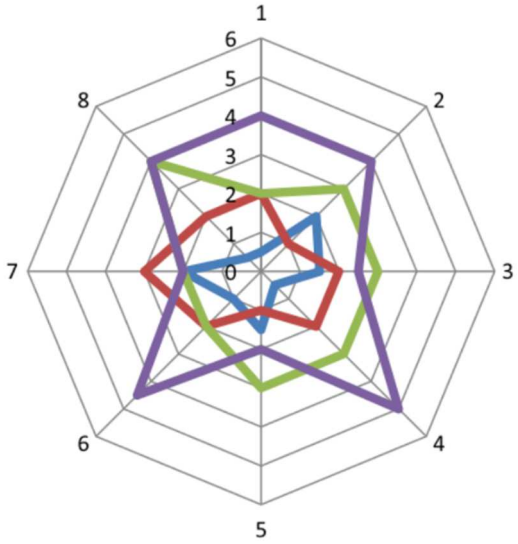

Figure 3: A schematic representation of the dimensional framework with eight dimensions (goal awareness, context awareness, space awareness, time awareness, bodily self-awareness, reflective self-awareness, group awareness, and workspace awareness). The coloured lines could represent four different kinds of agents or the awareness present at the local level vs Centralised levels for two agents (green-blue for agent 1; red-purple for agent 2).

4. Metrics

The measurement challenge is usually concerned with how we identify the presence (or absence) of awareness, or more precisely, determine the degree to which it is present, and identify some of its features. Defined as such, the goal of measurement assumes that awareness is a real property of the system and is less easily compatible with non-realist interpretations where attributions of awareness are supposed to capture a way to explain the system or interact with it. To avoid assuming a realist instance, we propose using the word "metrics" and defining it as a quantitative assessment of awareness for users' purposes.

4.1. Tests and metrics for dimensions

The first way to provide awareness metrics is to find tests for each dimension. Some tests can be adjusted from the psychological literature, including comparative psychology. For instance, to measure bodily self-awareness in infants or animals, psychologists have used the mirror test to assess the likelihood that the agent would recognise their own motion in a mirror and be able eventually to detect and act on a physical anomaly on their own body (e.g. a blue spot of paint placed on their face). Mirror recognition has been used to measure self-awareness in robots (e.g. Michel et al. 2004), and imitation tasks have been used to establish a self-other distinction (Suzuki et al., 2005)

4.2. Modulation of existing metrics

A second approach, which may be more directly relevant for industrial applications and users, considers the impact of awareness on existing metrics such as resilience, robustness or performance. To match with the dimensional approach, we can also represent the modulation on a dimensional radar plot and be used to compare and evaluate trade-offs between different scales of awareness.

Funded by the European Union
European Innovation Council
Funded by the European Union under Grant Agreement 101070918. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Innovation Council and SMEs Executive Agency (EISMEA). Neither the European Union nor the granting authority can be held responsible for them.

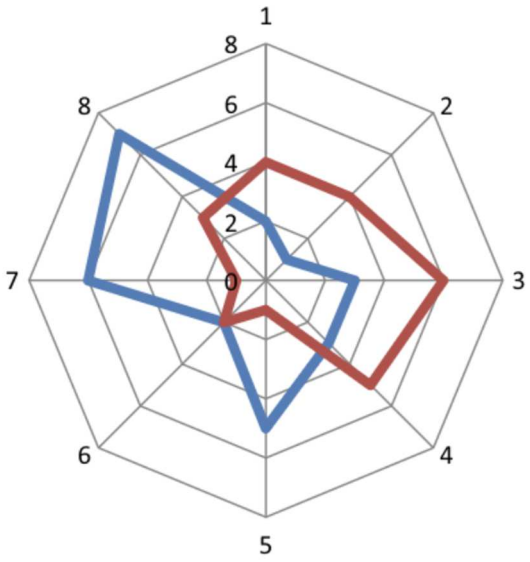

Figure 4: Schematic spider plot showing eight different metrics important for the use of a given agent (e.g. robustness, resilience, scalability, usability, trustworthiness, energy, cost, autonomy) and comparing a non-aware to an aware system (blue and red lines) or a more and less aware system.

5. Implementation

So far, the dimensional approach makes no assumption regarding the implementation of awareness. The next step will be to match dimensions of awareness to existing or new forms of architectures of awareness present in the literature. A large gap currently exists between the definitions of awareness and the architecture or implementation of awareness in artificial agents.

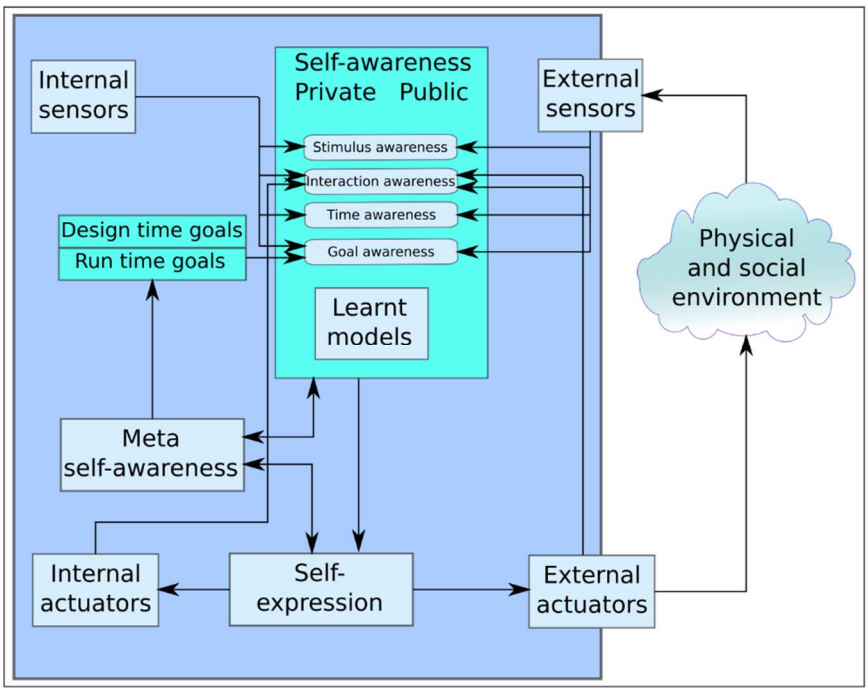

Figure 5: Architecture proposed for self-aware computing by Lewis et al. (2015).

Funded by the European Union
European Innovation Council
Funded by the European Union under Grant Agreement 101070918. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Innovation Council and SMEs Executive Agency (EISMEA). Neither the European Union nor the granting authority can be held responsible for them.

Frameworks like Lewis et al. (2015) (Figure 5 above) can be adjusted to list other dimensions of awareness but cannot capture the interdependence discussed above. Other accounts describing more systematic layering could capture the interdependence between different contents or kinds of awareness but would capture them through an unidimensional hierarchy, where one layer grounds the layer above. For instance, Lee et al. (2015) describe a five layers hierarchy architecture for self-awareness which is not mapped on contents or levels but distinguishes (1) a connection layer for sensors, (2) a conversion layer for parts with self-awareness, (3) a cyber layer for a machine-machine interaction (4) a cognition layer for self-prioritising and self-optimising decisions (5) a configure layer for self-optimisation, self-adjustment and self-configuration.

6. Ethics and human-AI interactions

Will the presence or degree of awareness present in AI change ethical issues? A straightforward question for artificial awareness is whether it carries the same ethical implications as considering "conscious AI" or "sentient AI". An entity's moral status has traditionally been linked to its level of consciousness. The presence or absence of consciousness is taken to be a crucial factor in determining whether some animals should be given moral consideration, such as in recent discussions on our treatment of cephalopods (e.g. Browning 2019) and crustaceans (Birch, 2017).

This moral status can dictate what actions are morally permissible, impermissible, or required when dealing with the entity. The entity's interests, preferences, plans, desires, or feelings should matter to some extent when determining its moral status. As far as the possession of certain of these attributes (e.g. feelings, desires) requires consciousness, the question of whether machines can possess consciousness is seen as intricately tied to the issue of their moral status (see, for instance, Gunkel, 2018; Metzinger, 2021; Mosakas, 2021). The core question is whether the ethical implications that have been discussed imagining that a machine could be conscious of the environment or have self-consciousness dissolve or remain once we consider a more austere framework where the system is aware of the environment or self-aware (see, for instance, the framing in Agar, 2019). For instance, mere awareness of stimuli without any accompanying 'feeling' that any experienced states or stimuli were good or bad may not trigger moral concerns about the possible experiences of this machine. This is, for instance, argued (although still using the word "consciousness") by Basl, 2014: "If we create a consciousness with only the capacity for experiencing colours, but with no attending emotional or other cognitive response, we need not worry about wronging said consciousness" (Basl, 2014, p. 84).

While this ethical issue remains to be addressed in more detail, two points can be highlighted. First, adopting a dimensional model of awareness may help avoid casting ethical debates as all-or-nothing, which the (highly speculative) presence of qualitative consciousness or subjective perspective in AI otherwise raises. A dimensional model can help, for instance, identify whether ethical issues depend on the system under consideration being more or less aware or on some specific dimensions of awareness, such as evaluative self-awareness.

It is essential to ask whether the ethical question only seems to matter if one adopts a realist reading of awareness - where awareness captures the real properties of the system. If awareness attributions are only part of the explanatory tools that we use to explain and predict the system's behaviour, the ethical implications seem to dissipate. Or do they? Normatively, as policies and ethics recommendations have to care about how users will relate to these

Funded by the European Union
European Innovation Council
Funded by the European Union under Grant Agreement 101070918. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Innovation Council and SMEs Executive Agency (EISMEA). Neither the European Union nor the granting authority can be held responsible for them.

technologies, they also need to consider and regulate the transactions between AI producers and users within the terms that are culturally or socially accepted. Coeckelbergh (2021), for instance, posits that societal norms and perceptions are more critical than ontology when it comes to ethics and policy surrounding AI.

An illustration coming here from the debates about trustworthy AI is that if users interacting with a chatbot or a caregiver robot are prone to trust it, then the law and ethics policies should use this category and see how the AI system meets this expectation. By analogy, if the system is interpreted as aware or presented as being aware to users, then awareness becomes a similar target of concern for ethical implications. The recommendation here is to provide a systematic mapping of the differences introduced in the user of a given system when the system is presented or interpreted as aware or not aware, or more or less aware.

Summary and conclusions

The benefits of approaching awareness dimensionally are outlined and allow us to provide a particularly appropriate framework to address conceptual, measurement, and ethical issues.

References

Agar, N. (2019). How to treat machines that might have minds. Philosophy & Technology, 33, 269–282.

Basl, J. (2014). Machines as moral patients we shouldn't care about (yet): The interests and welfare of current machines. Philosophy & Technology, 27(1), 79– 96.

Bayne, T., Hohwy, J., & Owen, A. M. (2016). Are there levels of consciousness?. Trends in cognitive sciences, 20(6), 405-413.

Benerecetti, M., Bouquet, P., & Bonifacio, M. (2001). Distributed context-aware systems. Human–Computer Interaction, 16(2–4), 213–228. https://doi.org/10.1207/S15327051HCI16234_06

Bermúdez, J. L. (2011). Bodily Awareness and Self-Consciousness. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199548019.003.0007

Birch, J. (2017). Animal sentience and the precautionary principle. Animal Sentience, 16(1).

Block, N. (2011). Perceptual consciousness overflows cognitive access. Trends in cognitive sciences, 15(12), 567–575.

Bongard, J., Zykov, V. & Lipson, H. (2006). Resilient machines through continuous self-modelling, Science, 314, 1118–1121.

Broadbent, E. (2017). Interactions with robots: The truths we reveal about ourselves. Annual review of psychology, 68, 627–652.

Brown, J. W. (1990). Psychology of time awareness. Brain and Cognition, 14(2), 144–164.https://doi.org/10.1016/0278-2626(90)90026-K

Funded by the European Union
European Innovation Council

Funded by the European Union under Grant Agreement 101070918. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Innovation Council and SMEs Executive Agency (EISMEA). Neither the European Union nor the granting authority can be held responsible for them.

Browning, H. (2019). What should we do about sheep? The role of intelligence in welfare considerations. Animal Sentience, 4(25), 23.

Bruno, M-A. et al. (2011) From unresponsive wakefulness to minimally conscious PLUS and functional locked-in syndromes: recent advances in our understanding of disorders of consciousness. J. Neurol. 258, 1373–1384].

Burton, J. W., Stein, M. K., & Jensen, T. B. (2020). A systematic review of algorithm aversion in augmented decision making. Journal of Behavioral Decision Making, 33(2), 220-239.

Campbell, J. (2007). What is the role of spatial awareness in the visual perception of objects? Mind & Language, 22(5), 548–562. https://doi.org/10.1111/j.1468-0017.2007.00320.x

Chaminade, T., Zecca, M., Blakemore, S. J., Takanishi, A., Frith, C. D., Micera, S., ... & Umiltà, M. A. (2010). Brain response to a humanoid robot in areas implicated in the perception of human emotional gestures. PLoS one, 5(7), e11577.

Chatila, R., Renaudo, E., Andries, M., Chavez-Garcia, R. O., Luce-Vayrac, P., Gottstein, R., ... & Khamassi, M. (2018). Toward self-aware robots. Frontiers in Robotics and AI, 5, 88.

Chou, W.-L., & Yeh, S.-L. (2012). Object-based attention occurs regardless of object awareness. Psychonomic Bulletin & Review, 19(2), 225–231.https://doi.org/10.3758/s13423-011-0207-5

Coeckelbergh, M. (2021). Time Machines: Artificial Intelligence, Process, and Narrative. Philosophy & Technology, 34(4), 1623-1638.

Dainton, B. (2013). The perception of time. In: Miller, K., & Dyke, Heather. A Companion to the Philosophy of Time (pp. 389–469). John Wiley & Sons, Ltd : Chichester, UK. https://doi.org/10.1002/9781118522097.ch21

De Graaf, M. M., & Malle, B. F. (2019, March). People's explanations of robot behavior subtly reveal mental state inferences. In 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI) (pp. 239-248). IEEE.

Dehaene, S., Lau, H., & Kouider, S. (2021). What is consciousness, and could machines have it? Robotics, AI, and Humanity: Science, Ethics, and Policy, 43-56.

Drury, J. L., Scholtz, J., & Yanco, H. A. (2003). Awareness in human-robot interactions. SMC'03 Conference Proceedings. 2003 IEEE International Conference on Systems, Man and Cybernetics. Conference Theme - System Security and Assurance (Cat. No.03CH37483), 1, 912–918. https://doi.org/10.1109/ICSMC.2003.1243931

Dutt, N., Jantsch, A., & Sarma, S. (2016). Toward smart embedded systems: A self-aware system-on-chip (soc) perspective. ACM Transactions on Embedded Computing Systems (TECS), 15(2), 1-27.

Endsley, M. R. (1988). Design and evaluation for situation awareness enhancement. Proceedings of the Human Factors Society Annual Meeting, 32(2), 97–101. https://doi.org/10.1177/154193128803200221

Endsley, M.R. (2013). Situation awareness. In: Lee, J. D., & Kirlik, A. (Eds.). (2013). The Oxford Handbook of Cognitive Engineering (Ser. Oxford library of psychology). Oxford University Press.

Funded by the European Union
European Innovation Council

Funded by the European Union under Grant Agreement 101070918. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Innovation Council and SMEs Executive Agency (EISMEA). Neither the European Union nor the granting authority can be held responsible for them.

Farina, L. (2022). Artificial Intelligence Systems, Responsibility and Agential Self-Awareness. In V. C. Müller (Ed.), Philosophy and Theory of Artificial Intelligence 2021 (Vol. 63, pp. 15–25). Springer International Publishing. https://doi.org/10.1007/978-3-031-09153-7_2

Gallagher, S. (2007). Bodily self-awareness and object perception. Theoria et Historia Scientiarum, 7(1), 53. https://doi.org/10.12775/ths.2003.004

Garner, W. R., & Felfoldy, G. L. (1970). Integrality of stimulus dimensions in various types of information processing. Cognitive psychology, 1(3), 225-241.

Grigore, E. C., Pereira, A., Zhou, I., Wang, D., & Scassellati, B. (2016). Talk to me: Verbal communication improves perceptions of friendship and social presence in human-robot interaction. In Intelligent Virtual Agents: 16th International Conference, IVA 2016, Los Angeles, CA, USA, September 20–23, 2016, Proceedings 16 (pp. 51-63). Springer International Publishing.

Ginsburg, S. & Jablonka, E. (2019). The Evolution of the Sensitive Soul: Learning and the Origins of Consciousness. Cambridge: MIT Press.

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. Science, 315(5812), 619-619.

Gunkel, D. J. (2018). Robot Rights. MIT Press.

Hmamouche, Y., Ochs, M., Prevot, L., & Thierry, C. (2020, February). Exploring the dependencies between behavioral and neuro-physiological time-series extracted from conversations between humans and artificial agents. In 9th International Conference on Pattern Recognition Applications and Methods (pp. 353–360). SCITEPRESS-Science and Technology Publications.

Hwang, S. (2019). A network clock model for time awareness in the Internet of things and artificial intelligence applications. The Journal of Supercomputing, 75(8), 4309–4328. https://doi.org/10.1007/s11227-019-02774-0

Jones, S., Milner, E., Sooriyabandara, M., & Hauert, S. (2020). Distributed situational awareness in robot swarms. Advanced Intelligent Systems, 2(11), 2000110.

Kahn Jr, P. H., Kanda, T., Ishiguro, H., Gill, B. T., Ruckert, J. H., Shen, S., ... & Severson, R. L. (2012, March). Do people hold a humanoid robot morally accountable for the harm it causes? In Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction (pp. 33-40).

Karpus, J., Krüger, A., Verba, J. T., Bahrami, B., & Deroy, O. (2021). Algorithm exploitation: Humans are keen to exploit benevolent AI. iScience, 24(6), 102679.

Kennedy, J., Baxter, P., & Belpaeme, T. (2017). Nonverbal immediacy as a characterisation of social behaviour for human–robot interaction. International Journal of Social Robotics, 9, 109–128.

Lee, M., Ruijten, P., Frank, L., de Kort, Y., & IJsselsteijn, W. (2021, May). People may punish but not blame robots. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (pp. 1-11).

Levy, D. (2009). The ethical treatment of artificially conscious robots. International Journal of Social Robotics, 1(3), 209–216.

Funded by the European Union
European Innovation Council

Funded by the European Union under Grant Agreement 101070918. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Innovation Council and SMEs Executive Agency (EISMEA). Neither the European Union nor the granting authority can be held responsible for them.

Lewis, P. R., Chandra, A., Faniyi, F., Glette, K., Chen, T., Bahsoon, R., ... & Yao, X. (2015). Architectural aspects of self-aware and self-expressive computing systems: From psychology to engineering. Computer, 48(8), 62–70.

Liu, H., & Wang, L. (2021). Collision-free human-robot collaboration based on context awareness. Robotics and Computer-Integrated Manufacturing, 67, 101997. https://doi.org/10.1016/j.rcim.2020.101997

Maniadakis, M., Trahanias, P., & Tani, J. (2009). Explorations on artificial time perception. Neural Networks, 22(5–6), 509–517. https://doi.org/10.1016/j.neunet.2009.06.045

Mahmoodi, A., Bahrami, B., & Mehring, C. (2018). Reciprocity of social influence. Nature communications, 9(1), 2474.

Metzinger, T. (2021). Artificial suffering: An argument for a global moratorium on synthetic phenomenology, Journal of Artificial Intelligence and Consciousness, Vol. 8, No. 1, 43–66.

Mosakas, K. (2021). On the moral status of social robots: considering the consciousness criterion, AI & Society 36, 429–443.

Nagel, T. (1974). What is it like to be a bat? The Philosophical Review, 83(4), 435–450.

Reggia, J. A., Huang, D. W., & Katz, G. (2015). Beliefs concerning the nature of consciousness. Journal of Consciousness Studies, 22(5-6), 146–171.

Rauchbauer, B., Nazarian, B., Bourhis, M., Ochs, M., Prévot, L., & Chaminade, T. (2019). Brain activity during reciprocal social interaction investigated using conversational robots as control condition. Philosophical Transactions of the Royal Society B, 374(1771), 20180033.

Salles, A., Evers, K., & Farisco, M. (2020). Anthropomorphism in AI. AJOB neuroscience, 11(2), 88-95.

Schwenkler, J. (2012). Does visual spatial awareness require visual awareness of space? Mind & Language, 27(3), 308–329. https://doi.org/10.1111/j.1468-0017.2012.01446.x

Signorelli, C. M. (2018). Can computers become conscious and overcome humans? Frontiers in Robotics and AI, 5, 121.

Suzuki, T., Inaba, K., & Takeno, J. (2005). Conscious robot that distinguishes between self and others and implements imitation behavior, Proceedings of the 18th international conference on Innovations in Applied Artificial Intelligence, 101–110.

Takiguchi, T., Mizunaga, A., & Takeno, J. (2013). A study of self-awareness in robots. International Journal of Machine Consciousness, 05(02), 145–164. https://doi.org/10.1142/S1793843013500030

Vuorio, R., Sun, S. H., Hu, H., & Lim, J. J. (2019). Multimodal model-agnostic meta-learning via task-aware modulation. Advances in neural information processing systems, 32.

Yanco, H. A., & Drury, J. (2004). "Where am i?" acquiring situation awareness using a remote robot platform. 2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No.04CH37583), 3, 2835–2840. https://doi.org/10.1109/ICSMC.2004.1400762

Zhao, S. (2018). What is reflective self-awareness for? Role expectation for situational collaboration in alliance animal society. Philosophical Psychology, 31(2), 187–209. https://doi.org/10.1080/09515089.2017.1392011

Funded by the European Union
European Innovation Council

Funded by the European Union under Grant Agreement 101070918. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Innovation Council and SMEs Executive Agency (EISMEA). Neither the European Union nor the granting authority can be held responsible for them.

Złotowski, J., Proudfoot, D., Yogeeswaran, K., & Bartneck, C. (2015). Anthropomorphism: opportunities and challenges in human–robot interaction. International journal of social robotics, 7, 347-360.

Funded by the European Union
European Innovation Council
Funded by the European Union under Grant Agreement 101070918. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Innovation Council and SMEs Executive Agency (EISMEA). Neither the European Union nor the granting authority can be held responsible for them.