EMERGE

WP1 Conceptual framework

D1.3 Dimensions of collaborative awareness

Version: 1.0

Date: 30/09/2024

European Innovation Council Funded by the European Union

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	D1.3 Dimensions of collaborative awareness
Contractual delivery date	30/09/2024
Actual delivery date	30/09/2024
Nature[1]	R
Dissemination level[2]	PU
Lead beneficiary	LMU
Editor(s)	Nadine Meertens (LMU)
Contributor(s)	Ophelia Deroy (LMU), Simon Jones (UOB), Jurgis Karpus (LMU), Bahador Bahrami (LMU)
Reviewer(s)	Sabine Hauert (UOB)
Document description	This report outlines a dimensional framework for understanding and measuring collaborative awareness, providing a conceptual toolkit for assessing collaborative behaviours in heterogeneous systems.

[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

European Innovation Council Funded by the European Union

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	Nadine Meertens	04.09.2024	TOC and sections outlined
0.2	Nadine Meertens	10.09.2024	Section 1 and 2 draft completed
0.3	Nadine Meertens, Ophelia Deroy.	12.09.2024	Section 7 draft completed, section 1 and 2 completed
0.4	Nadine Meertens, Ophelia Deroy, Simon Jones, Bahador Bahrami	18.09.2024	Section 3, 6 draft completed, Section 7 completed
0.5	Nadine Meertens, Jurgis Karpus	19.09.2024	Section 4, 5 draft completed, Section 6 and conclusion completed
0.6	Nadine Meertens	20.09.2024	Report draft completed and submission to reviewer
0.7	Sabine Hauert	26.09.2024	Review completed
0.8	Nadine Meertens, Simon Jones	29.09.2024	Integration of reviewer's comments
0.9	Nadine Meertens	30.09.2024	Final version of deliverable completed and submission to coordinator
1.0	Davide Bacciu	30.09.2024	Document submitted

[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.

European Innovation Council Funded by the European Union

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

Artificial systems are increasingly deployed in environments shared with humans and diverse artificial agents, making the ethical and effective collaboration between these entities a pressing concern. To address this, EMERGE advocates for the implementation of collaborative awareness in artificial systems. This approach extends beyond local awareness, highlighting the emergent group-level properties that support collaborative behaviour. By enhancing communicative abilities, these systems can achieve coordination and cooperation, crucial for working effectively within heterogeneous collectives. For both end-users and the public to trust and effectively utilise these systems, robust tools for describing and evaluating their collaborative capabilities are essential.

This report introduces a refined multidimensional framework for collaborative awareness in artificial systems, detailing potential measures for assessing and differentiating the collaborative abilities of various agents or collectives. It elaborates on the dimensions of awareness outlined in EMERGE, explores the formulation and comparison of awareness profiles, and integrates these concepts into a multidimensional awareness framework. Additionally, it discusses the roles of emergence and information sharing, culminating in a theoretical and conceptual toolkit for measuring and evaluating collaborative awareness, informed by insights from ethology, game theory, neuroscience, social psychology, and swarm robotics.

Consortium

The EMERGE consortium members are listed below.

Organization	Short name	Country
Università di Pisa	UNIPI	IT
TU Delft	TUD	NL
University of Bristol	UOB	UK
Ludwig Maximilian University of Munich	LMU	DE
Da Vinci Labs	DVL	FR

European Innovation Council Funded by the European Union

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

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

European Innovation Council Funded by the European Union

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 figures

List of tables

List of abbreviations

GA	Grant Agreement
CA	Consortium Agreement
IPR	Intellectual Property Rights
WP	Work Package

European Innovation Council Funded by the European Union

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.

Executive Summary

This document is a deliverable of the EMERGE project, funded under grant agreement number 101070918.

This deliverable, “D1.3 Dimensions of collaborative awareness”, reports on the activities of WP1 “Conceptual framework” which aims to elaborate and refine a new framework for collaborative awareness tailored for a set of heterogeneous agents.

In this document, we refine the multidimensional framework of awareness and integrate the concepts of collaborative awareness into it. Moreover, we explore existing approaches and measures to distinguish and assess collaborative behaviours from ethology and sociobiology, evolutionary game theory, neuroscience, social psychology, and swarm robotics.

1. Introduction

Consider a forest engulfed in a massive wildfire. To combat this unfolding natural disaster, fire departments and government agencies deploy numerous teams of heterogeneous robots. Some robots specialize in sensing and tracking the fire's movement, identifying the best spots to target. Others focus on coordinating the actions of various robots and human firefighters. Drones are soaring above, dropping water or fire-retardant chemicals to extinguish the blaze, while on the ground, autonomous intervention vehicles and human firefighters work side by side to contain the fire. This is a glimpse into the future of collaborative efforts between artificial agents and humans.

The development and exploration of such multi-agent, heterogeneous firefighting teams are well underway (Innocente and Grasso, 2019; McConville, 2024; Roldán-Gómez et al., 2021; Seraj et al., 2019; Tzoumas et al., 2023, 2024). Each robot in this scenario is designed for a specific task, built with its (potentially) unique architecture, and developed by a particular brand. Yet, for the team to function effectively, these robots must coordinate their efforts and collaborate both with each other and with human firefighters. Moreover, the success of these operations hinges on the human firefighters' ability to trust and rely upon the set of robotic systems—potentially with their lives, as well as the lives of the residents in the affected area.

This high-stakes scenario is here to highlight a much more common problem: the collaboration of multiple domain-specific systems – between themselves, with human collaborators, and with users - while operating in shared spaces.

Beyond the practical and technological hurdles of ensuring successful cooperation and coexistence, numerous ethical issues also emerge, particularly in high-stakes situations like the one described. To build trust and ensure the effective use and reliance on such systems by both end-users and the general public, it is not enough to simply deploy robotic teams; we must also develop robust tools to describe and evaluate them comprehensively – especially if we expect people to rely on them in high-stakes scenarios.

This entails two objectives.

First, to elaborate clear and precise language to avoid miscommunication and prevent misleading descriptions of these systems (Bones et al., 2021; Deroy, 2023; Dorsch and Deroy, 2024a; Dorsch and Deroy, 2024b).

Second, to provide tools for a rigorous assessment, which facilitates informed ethical use and effective policy regulations of these systems (Winfield, 2019).

European Innovation Council Funded by the European Union

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.

In previous deliverables and papers, EMERGE has addressed these two objectives at individual and collective scales. In D1.1, we introduced an operational concept of awareness, distinct from consciousness - with which it has often been conflated (see also Deroy et al., 2024). Integrating awareness into artificial systems allows for a shift away from centralised control, promoting local, autonomous, and adaptive functioning at both the individual and group levels. This approach can enhance efficiency, resilience, and flexibility (D1.1 local awareness criteria).

To facilitate collaboration among individual, potentially heterogeneous agents, we introduced the concept of collaborative awareness. This operational and ethically tractable concept captures the collaborative capacities that emerge at the group level (D1.2 demarcating collaborative awareness from related concepts). Our aim is to address the challenges of operating domain- or task-specific multi-agent heterogeneous systems in shared spaces – ensuring that these systems can be safely relied upon by human collaborators and users. Introducing collective awareness in systems of collaborating narrow agents is a solution to enable easier monitoring and interfacing between the artificial systems and human users (see Fig. 1 below).

This report tackles two key questions. First, how can we develop a multidimensional framework for understanding collaborative awareness? How do we define it, and how can we integrate it with the dimensional model of awareness introduced in D5.1, which focused on measuring emergent awareness? Second, what are the most effective methods for assessing the degree of awareness within collectives?

To answer these questions, we propose a theoretical and conceptual toolkit for measuring collaborative awareness. This toolkit operates a translation of methods from ethology, (evolutionary) game theory, neuroscience, and psychology. By addressing these questions, this report aims to not only deepen our understanding of collaborative awareness but also advance its practical application in multi-agent systems.

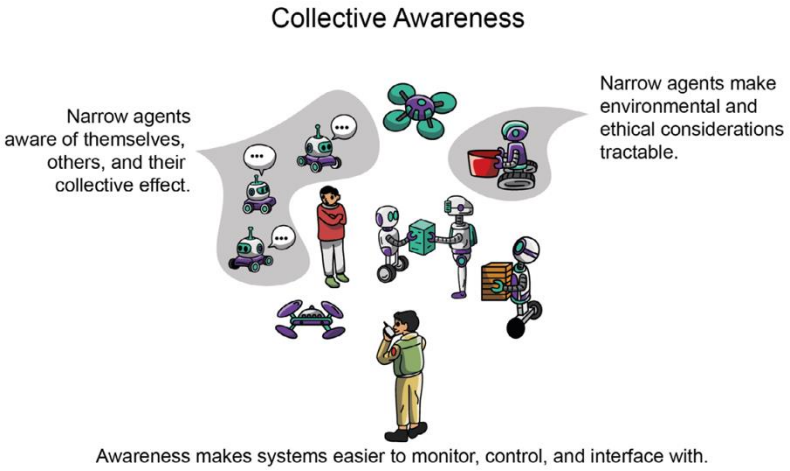

Figure 1: Features of collective awareness by Deroy et al. (2024).

2. A multidimensional framework for awareness

2.1 Dimensions of awareness

The notions of consciousness and awareness have often been used interchangeably, with awareness considered as part of a broader concept of consciousness, or awareness is

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.

employed in defining consciousness. While some distinctions have been drawn between the two, they remain unclear - particularly when it comes to operationalizing these concepts within artificial intelligence. The EMERGE project addresses this ambiguity by introducing a redefined concept of awareness, one that does not concern itself with the subjective experiences or 'inner lives' of artificial systems. Instead, it focuses on the capacities these systems can be given to adapt and navigate an ever-changing, complex environment. This new approach shifts attention away from what it might be like for an organism or system to experience something, and instead examines the abilities it has to interact with itself, its environment, and others - structured across various dimensions (Meertens, under review).

A dimensional approach was selected because it allows for a structured yet flexible investigation into the similarities and differences between systems, without imposing a rigid hierarchical structure or overlooking their uniqueness (see D1.1 Local awareness criteria). This multidimensional method stands apart from other approaches in cognitive science (Bayne et al., 2016) by offering a more fine-grained analysis that captures both the independence and interdependence of various aspects or capacities of awareness. For example, a swarm of drones and an artificial neural network (hereafter ANN) cannot be assessed in the exact same way - evaluation criteria designed for one may be misleading or entirely uninformative for the other. However, by adopting a more open, general, and additive approach, meaningful comparisons between such systems become possible. This mirrors existing approaches in philosophy of mind and ethology which engage in comparisons of cognitive abilities across species (Birch et al., 2020; Browning, 2023).

Building on the idea that awareness cannot be directly assessed but should instead be evaluated functionally - by observing the behaviour it enables - we propose associating each dimension of awareness with a range of capacities or abilities. These capacities are understood as properties of a system that contribute to its success in performing associated tasks, as measured by various performance metrics (see D5.1 Measuring emergent awareness). Performance, in this context, is seen as graded, highlighting the differences when a capacity is present versus absent. For example, in section 6, we will explore how spatial awareness can be evaluated in a swarm of robots by measuring their capacity to form a Distributed Reference Frame and how this improves performance in a swarm logistics task (see Jones and Hauert, 2024). This framework allows for the comparison and measurement of awareness across different artificial systems, though the precise dimensions to be considered remain an open question. Various dimensions of awareness discussed in the cognitive science literature were explored in D1.1; however, not all of these are consistent with one another, nor are they easily operationalized or suited for minimal systems that lack language capabilities. So, which dimensions might be worth exploring in greater depth, even if only tentatively?

Temporal awareness refers to an artificial system's ability to perceive, and act upon time-related aspects of its environment, its own actions, and others. This includes the capacity to recognize and respond to the timing of events, anticipate future states based on past experiences, and coordinate actions in relation to temporal factors. Temporal factors to consider here are: continuity, duration, simultaneity, persistence, change, succession, and an experience (flow) of past, present and future (Dainton, 2013; LePoidevin, 2019). In artificial systems, temporal awareness enables effective management of tasks that depend on timing, such as synchronising with other agents, predicting future occurrences, and adapting to changes over time.

In biological organisms, temporal perception plays a critical role in cognition (Varela, 1999), especially in conditioning and reinforcement learning (Gallistel and Gibbon, 2000). Temporal

European Innovation Council Funded by the European Union

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.

perception is also crucial for embodied and bio-inspired robotics (Maniadakis et al., 2009), as well as for multi-agent systems, particularly those involving human-robot interaction (Maniadakis et al., 2020) and heterogeneous teams of agents (Maniadakis et al., 2016). Research into temporal awareness in artificial systems has led to the development of mechanisms that enable systems to perceive and discriminate time durations. For instance, Maniadakis et al. (2009) evolved a Continuous Time Recurrent Neural Network (CTRNN) as a mechanism for time perception, applying it to a rule-switching task where a simulated agent adapted its behaviour based on changes in the environment, with varying durations across trials. Similarly, Lourenco et al. (2020) tested a simulated robot agent in a task inspired by Soares et al.'s (2016) research on mice, where the agent had to discriminate between shorter and longer time intervals. These studies demonstrate some possible ways of operationalizing temporal awareness in artificial systems, opening up opportunities for more advanced adaptive and responsive behaviours in dynamic environments. Implementing temporal awareness in artificial systems can take the shape of introducing a simple mechanism such as clock or standard time, to more complex mechanisms and abilities.

Spatial awareness refers to an artificial system's ability to perceive, process, and respond to its physical environment in relation to its own position and the positions of other entities or objects within that space. In artificial systems, this involves the capacity to navigate, localise, or coordinate movement or actions effectively by understanding spatial relationships. This ability enables systems to adapt dynamically to environmental changes, optimise task performance, and collaborate with other agents or humans in shared environments. Other capacities relevant to this dimension might involve sensing abilities to detect and interpret spatial features, obstacle avoidance, tracking others in the environment, awareness of proximity of others, as well as mapping the environment.

On a minimal level, artificial systems can rely on GPS for basic location tracking. Beyond this, robots could reply on feature-based localisation strategies, where the robot identifies and maps environmental features (e.g. Jones and Hauert, 2023). More advanced systems could rely on the capacity of Simultaneous Localisation and Mapping (SLAM). If an autonomous robot finds itself in an unfamiliar environment, with no existing GPS data available it needs to localise itself and work on constructing an incremental map of its environment – ideally at the same time (Saeedi et al., 2015). Saeedi et al (2015) point out how this only becomes more challenging when we consider robot teams, or an environment entailing multiple robots – especially if this concerns a distributed system (distributed SLAM, or DSLAM). However, such a distributed swarm will also provide more robust and potentially faster results (Birk and Carpin, 2006). The firefighting example from the introduction describes exactly such an application where SLAM or DSLAM would be a highly valuable capacity.

Metacognitive awareness refers to an artificial system's ability to monitor, assess, and regulate its own processes. This involves being aware of information it has available to it, decision-making processes, and problem-solving strategies, as well as the ability to adjust these processes when necessary. Metacognition construed as such is primarily self-directed. In artificial systems, metacognitive awareness could allow for self-evaluation, error detection, and the capacity to adapt strategies in real-time to improve performance or overcome challenges both individually and on a group level. On a minimal level this entails a capacity to provide confidence ratings for the decision or actions the system takes, and possibly an ability to opt-out of acting should this confidence rating be under a given (or learned) threshold. Another important concept in this context is **mentalizing** or **mindreading**, which involves the ability to engage with the mental states of others, making it inherently other-directed. The exact definitions of these terms, as well as their relationship, are subjects of significant debate (Proust, 2014). However, it is widely accepted that, for humans, the ability to monitor and

European Innovation Council Funded by the European Union

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.

respond to the mental states of others plays a crucial role in facilitating collaboration (Frith, 2012). In this field, a distinction is often made between explicit (verbal) and implicit (nonverbal) uncertainty monitoring. Additionally, in the study of animal cognition, there is an ongoing debate about whether observed behaviours should be interpreted as behaviour-reading or as evidence of (minimal) mindreading abilities (Buckner, 2014; Heyes, 2015; Lurz; 2015; Strasser, 2018).

Artificial systems are becoming increasingly complex, potentially giving rise to new problems in the collaboration or coexistence between heterogenous systems. However, this complexity raises even more pressing concerns when it comes to Human-Machine Interactions (HMI). Johnson (2022) suggests that implementing metacognition could be one was of addressing numerous safety and ethical concerns that arise. The starting point of such an approach can be located in implicit uncertainty monitoring and an ensuing fail-safe or opt-out mechanism. It would enable an AI system to prevent critical failures via self-diagnosis (Johnson, 2022). Moreover, implementation of decision confidence ratings or modulations in learning algorithms could vastly improve their efficiency and functioning in day-to-day decisions (Drugowitsch et al., 2019). Another existing implementation of metacognitive awareness in artificial systems is work of self-monitoring in autonomous systems (Mörwald et al., 2011; Chiba et al., 2020)

Agentive awareness has been discussed in various competing ways in cognitive science and philosophy, often referring to the phenomenology of agency, a sense of agency, the awareness of one's own actions, or the recognition of oneself as the agent currently acting (Bayne, 2011; Mylopoulos, 2017). In this view, agentive awareness is considered a specialized form of self-awareness (Bayne and Pacherie, 2007). These competing accounts primarily aim to capture a specific human experience, with phenomenal consciousness as a prerequisite. As a result, they have limited applicability to nonhuman animals or artificial systems.

To address this limitation, we propose that agentive awareness can be understood as having two layers: pre-reflective and reflective. **Reflective agentive awareness** aligns with traditional notions, where a system recognizes itself as the actor in a given situation, monitors its actions, and adjusts them if they do not align with its goals. This involves a higher-order evaluation of its performance and a capacity to correct errors. However, **pre-reflective agentive awareness** does not depend on self-reflection or explicit mental representation. Instead, it is a more immediate awareness of the world in terms of Gibsonian affordances (Gibson, 1979), where the system perceives the environment, itself, and others as presenting opportunities or constraints for action. This minimal form of pre-reflective agentive awareness allows a system to continuously evaluate its surroundings, dynamically adapt its behaviour, and interact with other agents based on how the environment and those agents afford certain actions. This enables the system not just to recognize objects or agents, but to perceive the potential actions they make possible, how those actions relate to its own goals, and how to influence or respond to the actions of others.

Much more can be said on this approach to agentive awareness, specifically the tension between representational accounts and Gibsons direct perception view (Chemero and Turvey, 2007). Despite this the ecological affordance-based component has already found popularity in robotics (Ardón et al., 2020; Horton et al., 2012).

Self-awareness in artificial systems, particularly its minimal instantiation of **bodily awareness**, refers to the system's ability to monitor its own physical states. In biological terms, this involves among other things body maps and proprioception, the system's capacity to perceive the position, movement, and orientation of its body or parts relative to its

European Innovation Council Funded by the European Union

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.

environment. Proprioception enables smooth, coordinated actions and self-regulation of movement.

Additionally, bodily self-awareness includes error or fault recognition, where the system detects discrepancies between its expected and actual physical state, such as malfunctions, or damage. In biological organisms the ability to detect noxious stimuli and bodily damage is referred to as nociception. For an artificial system such a capacity for detecting 'injury' might enable it to make real-time adjustments and signal the need for intervention. Beyond this passive monitoring, we can also distinguish more a more active approach. In cognitive science predictive processing theories explore how human beings (and possibly animals and AI) do not just interact passively with sensory input, but also actively predict sensory input before it arrives (Clark, 2015). This entails a pro-active engagement, where the system predicts its future state, is poised to act, and only then processes any potential deviations from the prediction. Understood in this sense bodily awareness need not be interpret as entailing only passive monitoring but could also built on active prediction.

Together, these capacities enable artificial systems to maintain effective and autonomous operation, especially in complex, dynamic environments. One example of such an application can be found in soft robots. Soter et al. (2018) implement bodily awareness through integration of exteroceptive and proprioceptive sensors. Their octopus-inspired arm uses four bend sensors in its soft body for proprioception, and a camera that records the movement of the arm for exteroception.

2.2 Integrating dimensions of awareness

This tentative description of the dimensions we aim to explore in EMERGE is not meant to be exclusive or all-encompassing. We want to point particularly at three different cases:

(1) **Missing dimensions**: other dimensions might be needed to capture abilities not addressed here.

(2) **Relation between dimensions**: the relations between these dimensions are yet to be determined, as they are likely to be related. Psychologists distinguish here between integral and independent dimensions. Integral dimensions often work together and are hard to separate because they form part of a unified experience. When you observe an object moving, for example, you are simultaneously aware of its position and the time over which the movement occurs. This suggests that spatial and temporal awareness go hand in hand. Independent dimensions, on the other hand, can function separately. For example, certain artificial neural networks (ANNs) have shown the ability to recognize the passage of time or duration without needing spatial awareness. This demonstrates that temporal awareness can exist independently of spatial capacities.

    Understanding when dimensions are integral or independent and for which system can help clarify how different instantiations of awareness operate within different systems, and how they may differ from biological or human cognition.

(3) **Emergence of new dimensions:** One area where the idea of emergence comes to light is with agentive and self-awareness, which come together is in the coordination of complex actions within shared environments. For both biological organisms and AI, this involves an understanding of the agent's own body and capabilities (self-awareness). However, it also entails the ability to initiate and control actions based on the awareness of one's own goals, goals of others in the shared space, potential

European Innovation Council Funded by the European Union

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.

shared goals, and input from the environment in terms of the potential actions it enables for different agents (agentive awareness). Awareness of peripersonal space can emerge from awareness of self and agentive awareness. Peripersonal space refers to the immediate area surrounding the body that an individual perceives as part of their personal space—essentially, the zone where objects or actions directly impact the body. For biological agents, this space is key to bodily awareness and survival, as it allows them to react quickly to threats or opportunities within their immediate environment. In humans, this involves complex neural processes that integrate sensory inputs, such as vision, touch, and proprioception, to create a real-time map of the body's location relative to external objects.

For AI and robotic systems, modelling peripersonal space is equally important. Autonomous agents need to understand their own boundaries in relation to their surroundings to act and interact effectively, whether avoiding obstacles or manipulating objects. Developing more sophisticated models of peripersonal space could significantly enhance the effectiveness of AI in complex, multi-agent interactions, helping them better navigate and cooperate in shared physical spaces.

2.3 Comparing awareness profiles

The evaluation of an artificial system on each of the dimensions, by associating them with capacities and associated tasks and metrics, can be merged into a dimensional model of awareness (see figure 2 and 3 below). Once enough data has been gathered along each dimension for a given artificial system a profile of awareness can be generated. This profile can then be compared to that of other similar artificial systems, or heterogenous (artificial) systems regarding both the overall "area" (quantity of awareness) and shape.

1. Comparisons of awareness profiles between similar systems are more feasible on the short term and gives rise to a more informative comparison. For instance, if we have three different soft robots that operate on a similar architecture their performance can be compared using the same experimental paradigms therefore generating a degree of awareness along each dimension between 0 and 1 which can fairly be compared. This could potentially generate the awareness profiles as illustrated in figure 2.

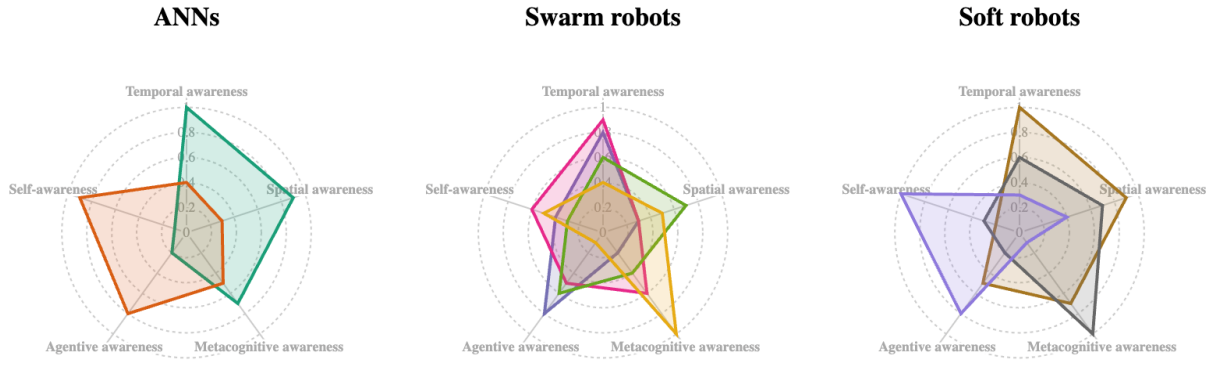

Figure 2: Hypothetical awareness profiles differentiated by artificial systems of different 'species'. Each spine of the graph represents an axis corresponding to a dimension of awareness, where the values along these axes reflect relative measures of the system's capability on a relative scale rather than in specific units.

2. Comparisons of heterogeneous (artificial) systems are possible but will either offer more general, less detailed insights or require the use of experimental paradigms typically suited for cross-species comparisons. Generating awareness profiles for such "interspecies" comparisons would result in figures like figure 3 (below).

European Innovation Council
Funded by the European Union

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.

The first route to constructing such a figure involves comparing systems either at the level of the task or at the level of the capacity.

For **task-based comparisons**, experimental paradigms broadly applicable across systems, as used in comparative cognition, can be utilized. Many tasks, like the mirror/mark test, duration tasks, and false belief tasks, have been applied to a variety of nonhuman animals, humans, and artificial systems.

For **capacity-based comparisons**, a detailed description of a target behaviour or ability is needed, along with criteria to assess the presence of this ability. These criteria could then result in various tasks that are formulated with a specific species in mind yet could be reasonably compared to other tasks aimed at the same capacity or behaviour.

For instance, Clayton and Dickinson's (1998) study on 'episodic-like memory' in scrub jays, which built on natural food-storing behaviour by the birds. The birds were trained with peanuts and wax worms. They could retrieve the food items after either 4h or 124h, the nuts would last until the 124h mark, but at that point the worms would be 'rotten' (made to taste unpleasant). The task for the birds then entailed remembering *what* was stored *where*, and *when* this occurred. Clayton and Dickinson picked out a target behaviour, namely episodic-like memory, and they selected criteria for this, namely recollection of 'what', 'where' and 'when'. These criteria have since been studied in other beings, including rats (Babb and Crystal, 2006), bottlenose dolphins (Davies et al., 2022), young children (Newscombe et al., 2014), and robots (Stachowicz and Kruijff, 2012).

The second route builds on the first but doesn't require a detailed awareness profile. Dimensional frameworks in consciousness (e.g., Birch et al., 2020; Dung and Newen, 2023) often use cross-species comparisons, although tasks may need to be adapted for different animals. By testing various abilities along the same dimensions—for instance, assessing spatial awareness in both an ANN and a soft robot—we can create awareness profiles for each. Even if the tasks aren't identical, we can gauge how well each system performs. For example, if the soft robot excels in 4 spatial awareness tasks while the ANN succeeds in only 1, we can infer stronger evidence of spatial awareness in the robot compared to the monkey. Though such comparisons are not fine-grained and lack precision, they can still be valuable, especially for ethical and policy considerations.

An example that captures both the challenges and ethical implications of cross-species comparisons can be located in work on pain capacity, this is the difficulty of translating pain research findings from animals, like rodents, to humans. As Cobianchi et al. (2014) explain, while rodent models are foundational to understanding pain mechanisms, there are often significant species differences that can complicate our understanding of how pain operates across different organisms.

Frameworks that have been developed for comparing welfare across different species in a scientific and ethical context, such as the mirror/mark test and duration tasks, offer a pathway forward. These tests, which have been applied to animals like chimpanzees, dolphins, and elephants, help assess higher cognitive abilities linked to awareness and potentially pain perception.

In situations where empirical data cannot conclusively establish the intensity of pain across species, assigning moral weights offers a practical alternative. This method allows decision-makers to assign relative importance to different species (or agents) based on ethical considerations, such as their capacity for welfare or the significance of their role within a given environment.

European Innovation Council Funded by the European Union

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.

For instance, a lion and a lungfish might have very different welfare capacities, but we could assign moral weight based on the complexity of their pain responses or cognitive functions (Browning, 2023). Similarly, an AI system that shows complex damage-avoidance behaviours might be assigned moral consideration proportional to its level of "awareness" or response complexity.

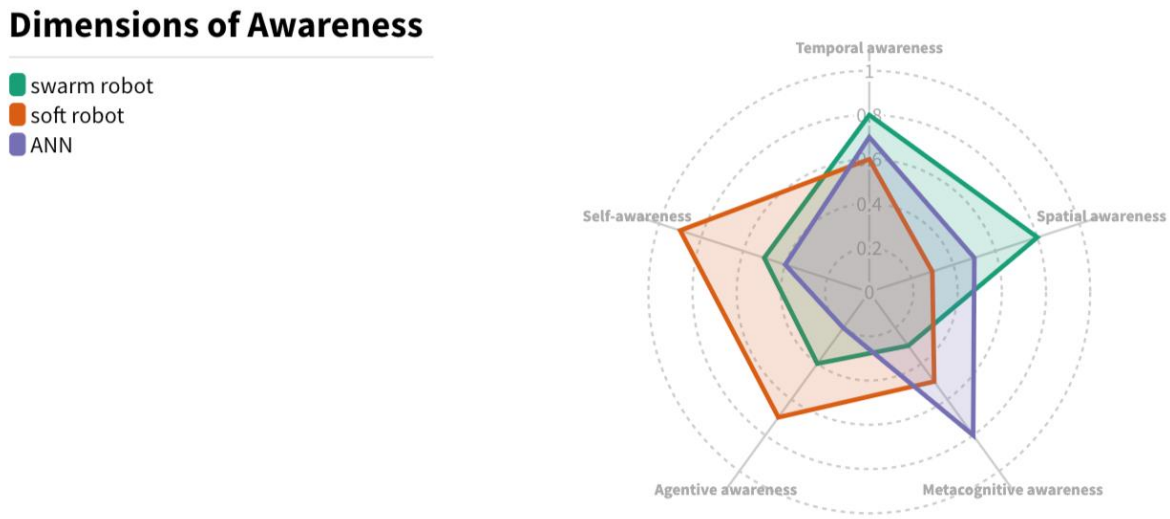

Figure 3: Hypothetical awareness profiles for swarm robots, soft robots, and artificial neural networks (ANN's). Each spine of the graph represents an axis corresponding to a dimension of awareness, where the values along these axes reflect relative measures of the system's performance or abilities in that dimension. The values are dimensionless, representing the system's capability on a relative scale rather than in specific units.

3. Collaborative awareness

3.1 Defining collaborative awareness

With the multidimensional framework of awareness established, we now turn to collaborative awareness. In D1.2, it was determined that existing concepts such as mutual awareness, collective awareness, and traditional notions of coordination and cooperation fall short in capturing collective behaviours and decision-making between artificial systems or between artificial systems and humans. To address this, we propose a tailored, conceptually engineered concept of collaborative awareness, consisting of three levels.

Collaborative awareness: the time-bound adaptive pursuit of multi-agent goals, including through changes in the environment and group rates or compositions.

Collaborative awareness, as defined here, serves as an umbrella term encompassing a wide range of collective behaviours and capacities. It spans from basic information exchanges to more complex forms of interaction, traditionally understood as collective or joint actions. A key aspect of this definition is the *adaptive pursuit* of shared or individual goals. This refers to the ability of systems or agents to dynamically adjust their behaviour in response to new information or changes in the environment, ensuring the successful achievement of a goal. In this context, a *goal* is an objective or desired outcome that drives the collaboration, which can be as straightforward as coordinating movements between robots to transport an object or as complex as solving a task through joint problem-solving between humans and AI. Such a goal can be, but crucially need not be, explicitly known to the agent. As long as its actions target a specific outcome adaptively, even if this aimed at outcome is predetermined and unchanging,

European Innovation Council
Funded by the European Union

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.

this suffices. The *time-bound* nature of collaboration, as approached in EMERGE, is crucial, as task completion is considered within a set or definitive time frame rather than open-ended. This aligns with the domain- or task-specific focus of the artificial systems at stake in this project. We are not targeting general systems that collaborate over long periods across a wide variety of tasks. Rather, we focus on systems designed for specific purposes, operating as efficiently and adaptively as possible within set parameters and objectives.

0 **Communication** (co-existence of goals)**:**
Communication is understood minimally in terms of the capacities to share information between agents, which can range from simple broadcasting or signalling to more sophisticated forms of interaction. At its most basic, this could involve the transmission of raw data or alerts, allowing systems or agents to be aware of each other's presence and status. At higher levels, it may involve the exchange of complex messages, where meaning is interpreted and context is understood, enabling more nuanced interactions. For example, two robots may share sensory data to inform each other of environmental conditions, even if they are pursuing independent tasks.

1 **Coordination** (interdependent goals)**:**
Coordination occurs when agents with distinct, yet interdependent, goals work together to ensure that their activities align for mutual benefit. Unlike communication, which merely involves information exchange, coordination requires agents to adjust their actions based on what others are doing, aiming to avoid conflicts or inefficiencies. This may include adapting timing, spatial positioning, or even sequences of tasks to ensure smooth interactions. For example, two drones performing different tasks in the same space may need to coordinate their flight paths to avoid collisions while accomplishing their respective objectives.

2 **Cooperation** (common goals):
Cooperation is the highest level of collaboration and occurs when agents share a common goal, leading to the alignment of their activities to achieve that shared objective. Here, the distinction between individual and group goals begins to blur, as the success of the group directly contributes to the success of each participant. Cooperation involves a more complex level of interaction, where agents may sacrifice individual benefits for the collective good. An example would be robots working together to assemble a product, where each action is dependent on the others, and the end goal (the completed product) is shared by all.

3.2 How do the levels combine?

In most scenarios, the levels of communication, coordination, and cooperation tend to form a hierarchical structure, where each builds upon the previous. Typically, communication serves as the foundation, enabling coordination by allowing agents to exchange information and adjust their actions accordingly. In turn, coordination lays the groundwork for cooperation, where agents align their activities toward a common goal. For instance, simple signalling at the communication level might be enough to facilitate basic coordination, even if more sophisticated forms of communication are absent. In this way, communication and coordination reinforce each other, providing the strategic framework for cooperation.

While the levels of communication, coordination, and cooperation might initially seem hierarchical, they are not strictly sequential. There are likely to be non-standard cases where this hierarchy does not strictly apply. Cooperation can occur without prior explicit communication, particularly in time-constrained situations. One way this happens is through

European Innovation Council Funded by the European Union

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.

pre-programmed or “instinctual strategies”. Similarly, coordination could be achieved through implicit mechanisms, such as shared environmental cues, rather than direct communication between agents. Another path to cooperation without communication is via established common ground or focal points (Schelling, 1960). For example, if a group of people, unfamiliar with each other, were told to meet in New York without the ability to communicate, studies have shown that many would converge on landmarks like Grand Central Station or Times Square around noon (Schelling, 1960; see also Mehta et al., 1994). These locations serve as salient focal points, guiding coordination without the need for explicit communication. Similarly, McMillan et al. (2012) examined how neural mechanisms support coordination in tasks without explicit communication. In a semantic task where participants are asked to name any boy's name, strategic decision-making leads them to consider what others are likely to choose, relying on social perspective-taking. This often results in common answers like 'John,' driven by probabilistic reasoning rather than direct communication (see also Bardsley et al., 2010; Isoni et al., 2013, 2019). While these cases are exceptions, they demonstrate that although the levels often interact and rely on each other, alternative pathways to coordination and cooperation may exist.

Moreover, in the levels can also interact by being instances of one another. When agents engage in **mutual communication**, they are typically coordinating their actions to some extent, even if it's just to ensure that the information is exchanged effectively. This requires them to adjust their actions—such as when to signal, how to interpret the signals, and how to respond—in response to the communication of the other agent. In this sense, communication itself often involves basic coordination because it requires a shared understanding of the rules or protocols governing the exchange of information (e.g., taking turns, using shared symbols, or agreeing on the medium of communication). The agents must adjust their behaviour in relation to each other to maintain effective communication, which is a type of coordination. However, not all forms of communication involve extensive coordination of broader actions or goals. For example, two systems might exchange signals without aligning their goals or coordinating further behaviours. But in more sophisticated forms of communication, especially when it involves complex exchanges (like negotiations or dialogue), coordination is inherently built into the process.

3.3 Dimensional model of collaborative awareness

Taking all this into account how might we formulate a dimensional model of awareness that also captures these levels of collaborative awareness?

The basic dimensional framework of awareness consists of five dimensions, each linked to various capacities that can be assessed through associated tasks and metrics. This framework aims to determine the degree to which an artificial system is aware along a given dimension. Metrics allow for the quantitative assessment of awareness for users’ purposes. This dimensional framework of awareness is captured in panel A in Figure 4 (below). Here we can distinguish between the directionality of the arrows. From a conceptual perspective, each dimension identifies the capacities of interest, which then inform the design of tasks and metrics. For end-users or developers assessing their systems, they will apply task-specific metrics to these tasks. By aggregating the performance across different tasks, they can assess the degree to which a system possesses certain capacities, which in turn when aggregated provides the degree to which that system is for instance, spatially aware. If this process is then repeated across all five (or potentially more) dimensions at stake, then a profile of awareness for this system is generated – such as presented in Figures 2 and 3 (above).

This awareness framework is versatile and can be applied at both local and global levels. At the local level, it can be used to assess a single artificial agent, evaluating its capacities and

European Innovation Council Funded by the European Union

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.

performance along the defined dimensions of awareness. For example, if we are examining a robot, we could measure its spatial awareness, temporal awareness, and other relevant dimensions using specific tasks and metrics. This process involves applying task-specific metrics to evaluate how well the robot performs in various tasks, such as navigating through an environment or coordinating movements. At the global level, the framework can be extended to assess entire swarms or groups of individual agents in much the same way.

However, that merely allows for an assessment and generation of the basic awareness profile for either individual or collective agents. How might we approach specifically looking at the collaborative abilities, and the ensuing awareness requirements?

Consider an individual agent within a distributed swarm. For this agent to effectively communicate, coordinate, or cooperate, it first needs to have information to share. If the agent possesses such information, such as spatial coordinates, this data must be in a format that is suitable for communication and intelligible to other agents. This entails that to communicate along the spatial dimension it needs a certain degree of spatial awareness. Collaborative awareness adds certain requirements or demand on to the awareness profile of an artificial system. When the agent needs to coordinate its actions with others or cooperate towards a common goal, the demands on its awareness increase. In this context, designing an artificial system with collaborative awareness necessitates specific requirements for capacities within the basic awareness profiles of the individual agents, or the collective as a whole. These requirements will vary depending on the domain, task, and the specific artificial system in use.

This interplay between awareness profiles and collaborative capacities can be examined from two perspectives. First, a specific awareness profile can enable a system to communicate, coordinate, and cooperate. Conversely, designing a system with the capability for collaboration necessitates having a certain awareness profile. This relationship is illustrated in Panel B of Figure 4 (below).

Instead of adopting a broad, domain-general perspective, we advocate for a domain-specific approach that focuses on the particular needs of the system. Deroy et al. (2024) highlight that domain-specific machines are more ethically manageable because they are easier to explain, control, and regulate. For effective collaboration among machines with different specializations or even different 'species' (i.e., brands or architectures), a concept of collaborative awareness is crucial. This framework should address varying demands for awareness within and across groups of collaborating systems.

To account for this, we propose a building-block or additive approach. This means that awareness is composed of distinct components or abilities, where different systems may possess some elements without necessarily having others. While certain capacities can build on one another, they don’t need to form a strict hierarchy. For example, a system might have advanced spatial awareness whilst only having limited metacognitive or self-awareness related abilities.

Such approaches offer several advantages and have been a significant topic in the consciousness and cognition literature (Spencer, 2024). This approach provides flexibility and customization by allowing researchers and developers to tailor assessments to specific needs, focusing on the capacities that are most relevant to their particular systems or applications. It supports a modular view of cognition, where complex phenomena are broken down into more manageable components, facilitating detailed analysis and incremental development. This approach aligns with the modularity of mind theories proposed in evolutionary psychology and cognitive science, which suggest that the mind is composed of specialized modules evolved

European Innovation Council Funded by the European Union

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.

to handle specific tasks (Spencer, 2024; Sperber and Mercier, 2018). Additionally, the building-block approach enables targeted research and practical application by allowing focused investigations into specific aspects of awareness or cognition, which is particularly useful for applied research and system development. This methodology also supports comparative analysis between systems with different specializations or architectures, helping to identify strengths and weaknesses in a systematic way.

To facilitate this, our framework supports a customizable approach where developers and end-users can select the capacities that are relevant to their specific system. Instead of a one-size-fits-all set of criteria for each dimension of awareness and level of collaboration, the base model provides a flexible structure that can be tailored. This allows users to define and operationalize the aspects of awareness that are pertinent to their systems, enabling targeted investigations and comparisons based on their specific needs and objectives. This personalized model ensures that the assessment of awareness aligns with the unique requirements of each application. Panel C in Figure 4 illustrates this by showing another potential add-on package of metrics (here ethics related ones), one could focus either the assessment or design of an artificial system on.

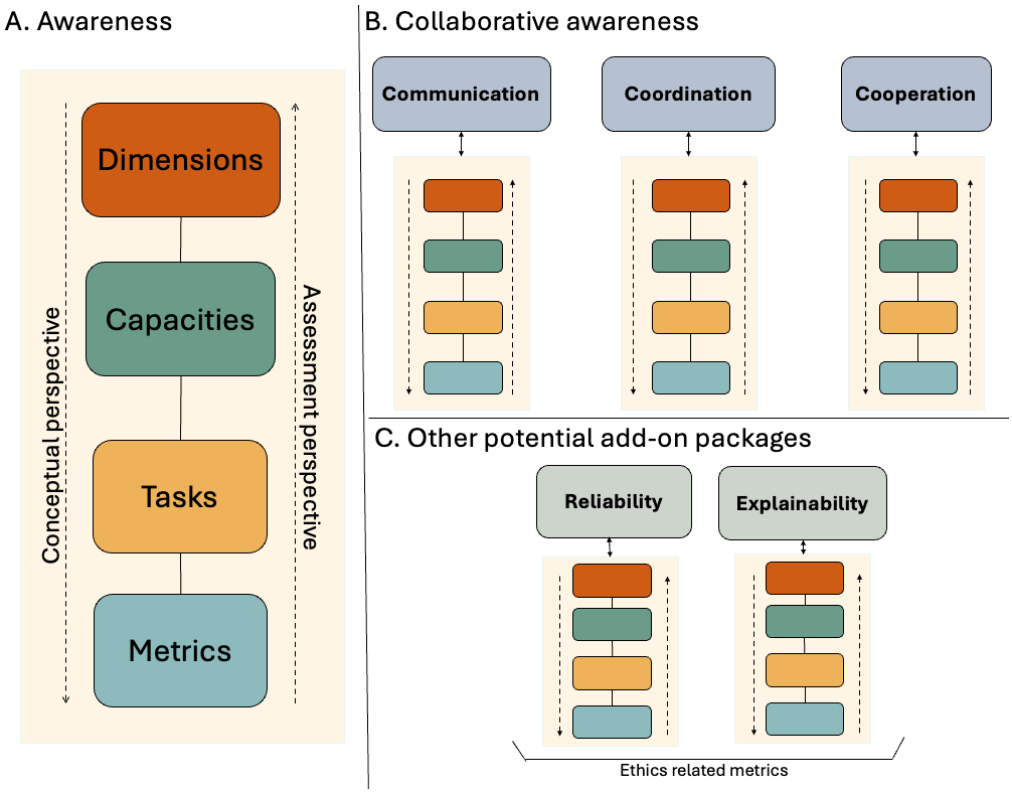

Figure 4: Conceptual framework of awareness in artificial systems. Panel A provides a schematic of the basic awareness model entailing dimensions, associated capacities, tasks and metrics. The arrows indicate the directionality either the conceptual or assessment perspective. Panel B and C are the same but for collaborative awareness and ethical awareness respectively.

4. Emergence

The dimensional framework of collaborative awareness presented in this report is grounded in the concept of *emergence*. But what does this entail? Specifically, how do collaborative behaviours emerge, and to what extent can properties like awareness be meaningfully applied at the group level?

European Innovation Council
Funded by the European Union

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.

Within the philosophical literature, the concept of emergence was first applied by Lewes (1877) in discussing the idea that in physical systems the whole is more than the sum of its parts (Clayton and Davies, 2006). In basic terms this means that at each level of complexity new qualities or properties emerge that cannot be attributed to the constituent parts. The example Davies gives is that water is justifiably described as wet, but it is meaningless to ask if the molecule H_2O is wet (in Clayton and Davies, 2006).

Usage of this term stands at a tension to the popularity of reduction and reductionist approaches in science and philosophy. Reductionism is the view that complex systems can be fully understood or explained by analysing their individual components and their interactions. In this view, the properties of a system are no more than the sum of its parts. For example, a car's ability to move can be reduced to the functioning of the engine, wheels, and other mechanical components without needing to invoke any additional concepts beyond those parts. The tension between emergence and reductionism lies in whether certain properties, such as consciousness, can be fully explained by their individual parts (as reductionism suggests) or whether new properties arise at higher levels of complexity that can't be predicted from just studying the parts (as emergence proposes).

Although this tension might seem insurmountable at a first glance, there are strong and weak version of both, each with their own advantages and disadvantages. For reductionism one can hold a **methodological reductionist** stance, where one approaches it as a useful method in science. Or one could take a position of **ontological reductionism** which holds a realist position regarding reduction.

Weak emergence holds that while higher-level properties or behaviours arise from the interactions of lower-level components, these emergent properties can still, in principle, be explained by the underlying rules governing simpler parts. However, in practice, predicting these outcomes from the base components can be too complex or impossible without direct observation or simulation. For instance, we might know the rules governing neurons, but the way a full brain produces conscious thought isn't something we can easily deduce just by studying individual neurons. Instead, we observe how consciousness emerges from these interactions as a higher-level phenomenon. Weak emergence suggests that these complex outcomes, while novel, are not separate or inexplicable in terms of their basic parts.

Strong emergence instead holds that micro-level principles are inadequate to account for the behaviour of a system as a whole. This means that there are properties of a system that arise from micro-level components but cannot be deduced (even in principle) from the principles that govern these micro-level components. Strongly emergent properties then cannot be reduced back to the fundamental parts of the system, examples often mentioned here are life and culture. In this sense, they are often described as being 'more than the sum of their parts'. For example, culture is considered irreducible, if one argues, it cannot be fully understood by examining individual behaviours in isolation. The interactions between individuals create complex social dynamics—such as shared beliefs, norms, and practices—that arise at the collective level and cannot be predicted by merely aggregating individual actions.

The positions of ontological reductionism and strong emergence are incompatible, but weak emergence can be combined with strong reductionism – and strong emergence with weak reductionism. A contentious debate rages on these positions. For the purposes of this report and EMERGE at large, it suffices to say that although a position of strong emergence could be argued for and defended, a commitment to **weak emergence** is already sufficient for our purposes. Weak emergence allows for the explanation of collaborative behaviours and awareness in artificial systems without requiring fundamentally new principles beyond the lower-level rules governing individual agents. We can ask and explore how awareness and

European Innovation Council Funded by the European Union

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.

collaboration emerge from interactions between agents, and how new behaviours emerge at the group level, but we do not need to propose that these higher-level phenomena are irreducible to their basic components *in principle*. Instead, we recognize that while these higher-level behaviours could potentially be traced back to the underlying processes of individual systems, the complexity is such that, practically, we need direct observation, simulation, or empirical study at the macro-level to understand them fully.

Such a position is desirable as it does not entirely rule out a position of strong reductionism. Essential elements to consider for the awareness framework as presented here are: non-aggregativity and distinctive efficacy. These entail that emergent properties, such as those described in the previous section on collaborative awareness, manifest in novel behaviours that are (in principle) not possible without them. Moreover, emergent collaborative dimensions yield unique abilities that are not in principle reducible to the local dimensions, due to their complexity.

5. Shared information in collective systems

In the context of communication, coordination, and cooperation among artificial systems, the concept of "information" plays a foundational role. In section 3 of this report, communication was defined as a capacity for sharing information between agents. But what exactly is meant by "information" in this context?

Although the use of an information-based approach to communication has been controversial in the study of animal signalling (see Seyfarth and Cheney, 2003), it is foundational in artificial intelligence (Floridi, 2011). In everyday language “information” is often used interchangeably with data, text, or any content transmitted through any medium. However, from a technical standpoint, information involves two key elements: it is extensive, and it reduces uncertainty (Adriaans, 2024). This means that information is additive, and if one were to have the complete information on a topic, their uncertainty would be reduced to zero.

In artificial systems, to generate collaborative awareness, information sharing is crucial. Whether it takes the form of sensory inputs—collected through cameras or sensors—or more abstract symbolic representations like maps or action plans, the sharing and interpretation of information underpins interactions between agents. Importantly, information need not be transmitted directly. Changes in the environment, such as altered positioning or movement patterns, can also serve as indirect communication. These actions function as signals that, when observed, provide critical information to other agents.

For example, consider a distributed robot swarm, with each agent possessing some sensory input with varying degrees of certainty. These robots can share their information in several ways: through direct communication channels like signals (e.g. flashing lights), or through indirect means such as leaving pheromone-like trails or modifying their movements. A robot encountering an error might signal its status by moving to the periphery of the swarm, providing valuable information to others in the group. What matters is not only that the information is transmitted, but that it is received, interpreted, and acted upon, leading to a change in the behaviour of the recipient.

This brings us to an essential aspect of collaborative awareness in artificial systems: **uptake** and **causality.** For communication to take place and be observable from an external perspective, it must alter the behaviour of the receiving agent in a meaningful way. If the information has no causal effect, it cannot be considered to have been successfully communicated. It must lead to some updating-process on the part of the receiver.

European Innovation Council Funded by the European Union

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.

In artificial systems, information functions as more than just the medium of interaction—it is the foundation for emergent collective behaviour. The type, structure, and flow of information between agents directly shape the system's ability to communicate, coordinate, and cooperate. Systems that can process and transmit information in ways that enable mutual understanding, and the formation of collective goals exhibit advanced forms of collaborative awareness. This involves more than simply transmitting raw data; it requires building shared models of the environment and aligning individual goals to collective action. In this sense, information becomes the mechanism through which distributed agents become aware of each other's intentions, plans, and needs, enabling collaborative awareness that surpasses the action-perception abilities of any single agent.

In the context of AI, this information-driven approach offers the possibility of true collective intelligence, where the whole system's capabilities exceed the sum of its parts. The challenge lies in designing systems that not only exchange data, but do so in ways that foster communication, coordination, and cooperation at a sophisticated level.

6. Measuring collaborative awareness: a theoretical toolkit

To establish effective methods for assessing the degrees of collaborative awareness in collectives of artificial systems, we need to understand the target phenomena. This entails formulating descriptions of what communication, coordination or collaboration look like in diverse systems, developing criteria for recognising (or designing) these abilities and possible experimental paradigms that might be relied upon. To make a start at this we explore established measures and concepts from ethology, evolutionary game theory, neuroscience, and social psychology.

6.1 Ethology and sociobiology

Drawing on the literature on collaborative behaviours in nonhuman animals (hereafter animals), we rely on insights from ethology, comparative psychology, and sociobiology. While a comprehensive overview of this body of research is beyond the scope of this report, we will discuss some key measures, descriptions, and criteria relevant to the target phenomena.

Starting with **communication**, Wilson (2000) defines biological communication as "the action on the part of one organism (or cell) that alters the probability pattern of behaviour in another organism (or cell) in a fashion adaptive to either one or both of the participants (p. 176)." In this sense, communication involves both a signal and a response, where the key indicator is a change in behaviour resulting from the communicative act—what can be described as 'uptake.' This emphasis is primarily practical. For humans, it's easy to imagine receiving new information and storing it without altering behaviour. In such cases, we can verify understanding through verbal report. However, with animals that lack this capacity for verbal feedback, we must rely on observable behavioural changes to confirm uptake.

Wilson also points out that not all actions that influence another's behaviour qualify as communication. For example, an attack will most certainly change an animal's behaviour but is hardly communicative. Additionally, for an action to count as communication, it must be sufficiently consequential. One animal pausing to watch another walk by, for instance, doesn't constitute communication. Such concerns could also by addressed by emphasising that the behaviour of the caller is goal-directed, however this comes with its challenges.

European Innovation Council Funded by the European Union

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.

Seyfarth and Cheney (2003) give the example of a male frog giving an advertising call to attract females who respond by producing one of two sounds, meanwhile also attracting a predating bat. The frog evolved with the goal of communicating its size and condition to both rivals and mates, whilst giving as little as possible information to the bats. This example indicates the limits of insisting on goal-directedness, under no reading is the frog 'intentionally' communicating with the bat, it goes counter to its goals to do so. Yet, it is communicating with the bat, as an unintended consequence of its communication with mates. Such evolutionary trade-offs highlight two points according to Seyfarth and Cheney. Namely, that communication is a social event, designed to impact listeners, and that the function of a signal can be asymmetric between listener and signaller.

Overall, there is no clear consensus on what constitutes communication in animals. Historically, the debate has centred on two major approaches: information transmission versus causal influence (Kalkman, 2019). Kalkman discusses a hybrid approach, which emphasizes causal influence mediated via information transmission, though it risks being too inclusive by counting co-adapted interactions as communication.

At the minimal end of the communication spectrum, we can differentiate between *broadcasting* and *signalling*. *Broadcastin*g involves sending information to a broader, non-targeted audience, whereas *signalling* is directed at specific recipients or groups. For example, mating calls by male frogs or predator alert calls by vervet monkeys (Seyfarth et al., 1980) are meant for a particular audience and thus fall under signalling. In contrast, the honeybee's waggle dance (von Frisch, 1954) is an example of broadcasting. Performed by foraging bees, the dance conveys information about the direction and distance of a food source to any observer. As Wilson (2000) notes, this form of communication is genetically fixed, with a one-to-one correspondence between the dance and its meaning, and it assumes an appropriate audience is present. Unlike the rigid rules of the waggle dance, vervet monkeys' alarm calls are more context-sensitive and adaptable to the environment (Deshpande et al., 2023).

In contrast to the indiscriminate broadcast alarm calls found in many monkeys and lesser apes, great apes exhibit more complex forms of communication, including vocalizations that appear to be used in goal-directed, intentional ways. Studies by Crockford et al. (2012) and Schel et al. (2013) have shown that apes, such as chimpanzees, use vocalizations not simply as alarm calls but with the intent to influence the behaviour of specific individuals. This behaviour aligns with broader evidence of intentionality in ape communication. Byrne et al. (2017) note that apes regularly engage in audience-targeting behaviours, waiting for responses and persisting or elaborating their signals when the intended recipients fail to react. Such examples of purposeful signalling suggest that apes can adjust their communication based on social contexts, making their vocalizations more sophisticated.

As discussed in D1.2, the existing literature on **coordination** and **cooperation** in animals suffers from a lack of clear or agreed-upon terminology, with terms like "cooperation" and "joint action" often used interchangeably. This terminological ambiguity reflects a deeper issue in understanding the complexity of these behaviours. Coordination, where individuals align their actions to achieve separate goals, is well-documented across species, but true cooperation—where agents share a joint goal and act toward its fulfilment—is only present in a minimal sense (if at all).

Depending on one's definition of cooperation, there are high levels of cooperation in some insect societies, specifically in eusocial insects. This can be explained through the indirect fitness benefits for each member of the colony due to their high relatedness with one another (Hamilton, 1964). The 'goals' of each member of such a colony are entirely unified with that of

European Innovation Council Funded by the European Union

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.

the group, moreover there is evolutionary pressure to prioritise the wellbeing of the group over that of any individual. Cooperation in this sense, between unrelated individuals is rare among nonhuman animals, and highly debated (Clutton-Brock, 2009).

However, in most animals, awareness is fundamentally individualistic. Even in instances of sophisticated communication and coordination, such as those observed in great apes, the awareness guiding behaviour remains tied to individual intentions and perceptions. Great apes, for example, are highly flexible in their collaborative behaviours, monitoring where conspecifics' attention is fixed, and drawing this attention to themselves in various ways (Tomasello, 2022). However, as Tomasello notes, they lack joint attention and cannot form mutually obligating joint goals or commitments. While they can coordinate their actions, such as in tasks that require timing or synchronised efforts (Visco-Comandini et al., 2015; Voinov et al., 2020), they do so without the kind of shared intentionality that underlies human cooperation (Call, 2009).

This limitation is key when distinguishing between coordination and cooperation. In coordination, each individual has its own goal, and their alignment of actions is largely pragmatic, often driven by individual gain rather than collective purpose. However, this does not mean the individual goal cannot by circumstance (or through evolution) overlap with the collective goal - it just means it is currently not shaped by it. Cooperation, however, involves agents committing to a shared goal, with their actions driven by a collective understanding of what they are jointly trying to achieve. This requires the presence of collaborative interactions where participants share psychological states with another - What Tomasello and Carpenter (2007) refer to as 'shared intentionality'.

The transition from coordination to cooperation may rely on cognitive mechanisms like recursive intentions—where an individual not only recognizes the goal of another but also understands that the other recognizes their own goal. However, this level of recursive thinking appears to be largely absent in animals.

In artificial systems, however, the potential for true cooperation takes on a different form. Unlike biological systems, AI can be designed with the capacity to "share" a model for collective action. This introduces a fundamentally different framework, where agents are not just aligned in their actions, but actively share a model of the world, task, or goal, making cooperative behaviour not only possible, but different from humans.

Table 1: overview of existing measures of collaborative awareness in animals

Measures of collaboration in Animals		
Communication	Signal diversity	The variety of communication signals used by a species, including vocal, visual, chemical, and tactile cues
	Signal complexity	The intricacy of individual signals, such as the structure of vocalisations or the elaborateness of visual displays.
	Contextual flexibility	The ability to use different signals in various contexts or modify signals based on the social environment.
	Individual distinctiveness	The presence of individually distinct signals, such as the signature whistles in bottlenose dolphins.
	Information content	The amount and type of information conveyed through signals, such as identity, emotional state, or intentions.
	Signal honesty	The reliability of signals in conveying accurate information about the sender's quality or state.

European Innovation Council Funded by the European Union

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.

Coordination	Synchronisation	The ability of individuals to align their behaviours or movements in time and space.
	Role recognition	The capacity to understand and respond to a partner's role in a cooperative task.
	Temporal coordination	The timing and sequencing of actions between individuals during collaborative efforts.
	Spatial coordination	The ability to maintain appropriate positioning relative to others during group activities.
	Task division	The allocation of different roles or subtasks among individuals in a coordinated effort.
	Behavioural matching	The degree to which individuals mirror or complement each other's actions.
Cooperation	Succes rate	The frequency with which individuals successfully complete cooperative tasks.
	Efficiency	The speed or resource utilisation in achieving cooperative goals.
	Reciprocity	The extent to which cooperative behaviours are reciprocated among individuals.
	Coalition formation	The ability to form alliances or partnerships for mutual benefit.
	Resource sharing	The willingness to share food, information, or other resources with conspecifics.
	Altruistic behaviours	Actions that benefit others at a cost to the individual performing them.
	Conflict resolution	The ability to resolve disputes and maintain cooperative relationships.

6.2 Evolutionary game theory

In this section, we turn to evolutionary game theory as a framework for understanding communication, coordination, and cooperation in artificial systems. Evolutionary game theory provides a model for how strategies and behaviours can evolve over time through interaction, competition, and adaptation, offering insights into the emergence of collective behaviours in both biological and artificial systems.

In game theory, a game refers to an interaction between two or more agents (players), where the outcome for each participant is determined not only by their own choices but also by the strategies chosen by the other players (Maynard Smith, 1982). These players can be any type of interactive decision-making agent, animals, plants, microorganisms, and particularly as of late also robots and simulated artificial agents (Hummert et al., 2014). A specific focus in early developments of evolutionary game theory has been on conflicts among animals within and across species, such as combat between two males of the same species for a mate or resources (Maynard Smith and Price, 1973).

A key point of contention is whether or not **communication** is by necessity at the benefit of both the actor and reactor and how such behaviours could evolve for both interspecies and intraspecies scenarios. In evolutionary terms, signalling must offer some adaptive benefit to both parties to be stable (Skyrms, 2010).

The challenge for **coordination** lies in avoiding misaligned strategies that lead to suboptimal outcomes for the group. This is particularly relevant in coordination games where individuals must choose actions that depend on predicting the choices of others, such as in mate selection or foraging behaviour (Skyrms, 2004). The Nash equilibrium selection problem (another name for the coordination problem in game theory) has led to the development of a menagerie of theories to explain how interactive agents manage to solve it in various domains: focal points among outcomes (Schelling, 1960), refined strategies of players (Harsanyi and Selten, 1988),

Funded by the European Union

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.

social norms (Binmore, 2005; Bicchieri, 2006), team reasoning (Sugden, 1993; Bacharach, 2006; Karpus and Radzvilas, 2021), and others.

Regarding **cooperative** behaviours, Axelrod and Hamilton (1981) point out how most of adaptation in biological evolution may have been assigned to selection at the level of populations, or entire species of organisms If that is true, cooperation among individual members of a population is quite likely too. However, if we focus on adaptation at the level of an individual interesting questions emerge in accounting for cooperative and altruistic behaviours on both intraspecies as well as interspecies levels. Within a single species, cooperation and altruism are both adaptive if there is a sufficiently close relation between players – kinship (Hamilton, 1964; Maynard Smith, 1964). However, cooperative and altruistic actions also take place when there are no kinship relations, such as in instances of mutualistic symbioses and in a large variety of interactions among humans (e.g. Camerer, 2003; Falk et al., 2003). The emergence of cooperation has also been observed in artificial systems like robotics, where experiments show that group-level selection and high relatedness between robots can promote cooperation, as seen in the evolution of communication in robotic foraging tasks (Hauert et al., 2014).

From the point of view of orthodox evolutionary game theory, cooperative and altruistic behaviours can emerge only when there is either a direct or indirect fitness increase to the individual interacting parties (Melis and Semmann, 2010), though alternative views have been proposed in game theory more widely as of late (e.g. Sugden, 1993; Bacharach, 2006; Radzvilas and Karpus, 2021). Cooperation, a game-theoretic understanding of it, often involves making some form of investment: engaging in behaviour that reduces the immediate payoff of the actor while increasing that of another party, but if enacted reciprocally is mutually beneficial to all reciprocally cooperating parties involved. As Bshary and Raihani (2017) point out, such behaviour is particularly vulnerable to cheaters and, hence high relatedness (kin or other) between individuals can safeguard one from this.

Table 2: overview of existing measures of collaborative awareness from game theory.

Table 2: Measures of collaborative awareness from game theory		
Communication	Emergence of signalling techniques	Spontaneous emergence of signalling techniques between interacting parties despite the possibility of mis-coordinated interaction of those signals between senders and receivers in light of possible deceit.
Coordination	Convergence on equilibria	Convergence of the interacting agents' actions on equilibria in the face of Nash equilibrium selection problems in repeated interactions.
Cooperation	Attainment of mutually beneficial outcomes	Achieving outcomes that are good for all parties involved as a group, but not necessarily for each individual party in isolation – in one-shot or repeated interactions in mixed-motive games, such as the Prisoner's Dilemma, Chicken, and Trust (e.g. the emergence of tit-for-tat reciprocation of cooperative and non-cooperative actions in the latter case).

6.3 Neuroscience

From a neuroscience perspective, communication, coordination, and cooperation are interrelated processes that involve complex neural mechanisms. Communication relies heavily on language-related brain regions like Broca's and Wernicke's areas, as well as the superior temporal gyrus for speech processing. Coordination engages areas such as the prefrontal cortex and temporoparietal junction, which are crucial for synchronizing actions and understanding others' intentions. Cooperation activates reward centres like the ventral striatum and regions associated with social cognition, including the medial prefrontal cortex.

European Innovation Council Funded by the European Union

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.

One of the most investigated collaborative skills - theory of mind - comes from the ability to attribute mental states, beliefs, intentions, and desires to oneself and others. Neuroimaging studies have consistently implicated several key brain regions in ToM tasks, including the medial prefrontal cortex (mPFC), temporoparietal junction (TPJ), superior temporal sulcus (STS), and precuneus. The mPFC is thought to be involved in reasoning about mental states, while the TPJ plays a role in distinguishing between self and other perspectives. The STS is associated with processing social cues and biological motion.

Joint action, which involves coordinating one's actions with others to achieve a shared goal, relies on many of the same neural substrates as ToM but also engages additional regions. Studies on the neural correlates of joint action have highlighted the importance of the mirror neuron system, including the inferior frontal gyrus and inferior parietal lobule. These regions are activated both when performing an action and when observing others perform similar actions, facilitating action understanding and imitation.

Additionally, joint action tasks often show increased activation in areas associated with executive function and cognitive control, such as the dorsolateral prefrontal cortex and anterior cingulate cortex. These regions likely support the coordination and monitoring of shared task goals.

Table 3: overview of existing measures of collaborative awareness from neuroscience.

Measures of collaborative awareness from neuroscience		
Communication	Interbrain synchronisation	Measuring neural coupling between individuals, especially in prefrontal and temporoparietal regions, during communicative tasks using techniques like fNIRS hyperscanning.
Coordination	Motor synchronisation	Assessing the timing and precision of coordinated movements between individuals, often measured through motion capture or EMG.
	Predictive motor processes	Measuring anticipatory adjustments in motor planning areas like the premotor cortex and cerebellum during joint tasks.
	Shared representation formation	Assessing activity in regions associated with action understanding and planning during coordinated tasks.
Cooperation	Reward system activation	Measuring activity in the ventral striatum and orbitofrontal cortex during cooperative versus competitive tasks.
	Trust-related neural responses	Assessing activity in regions like the anterior cingulate cortex and insula during trust-based cooperative interactions.
	Theory of mind engagement	Evaluating activity in the medial prefrontal cortex and temporoparietal junction during perspective-taking in cooperative contexts.
	Prosocial decision-making	Measuring activity in regions like the temporoparietal junction during choices that benefit others or the group.

6.4 Social psychology

While the definitions of communication, coordination and cooperation applied in social psychology are consistent with those applied to animals, and in neuroscience, a key dimension added by social psychology has to do with social norms.

Social norms are unwritten rules and shared expectations that dictate appropriate conduct in specific social contexts. They serve as informal guidelines that help maintain order, facilitate social interactions, and promote cooperation among group members. Norms can vary widely across cultures and subgroups, influencing everything from daily etiquette to more significant

European Innovation Council Funded by the European Union

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.

social behaviours. By providing informal yet powerful guidelines, social norms help to maintain order and predictability in interactions, thereby reducing the cognitive load on individuals by offering a clear structure of acceptable behaviours.

Bicchieri's (2017) theory of social norms offers a valuable framework for understanding how these norms emerge, how they guide behavior, and how they influence social dynamics. According to Bicchieri, social norms are not merely internalized beliefs but conditional preferences that depend on individuals' expectations about the behavior of others. She differentiates between two types of expectations:

- empirical expectations, which refer to what individuals believe others will do,
- normative expectations, which refer to what individuals believe others think they ought to do.

For a social norm to be followed, individuals need both types of expectations to align: they must expect that others will behave according to the norm and that others expect them to do the same. A key element of her framework is the idea that people comply with norms when they believe that most others are also complying (empirical expectation), and when they believe they are being observed or judged by others (normative expectation). This creates a feedback loop where individuals' behavior reinforces the collective adherence to the norm, fostering coordination and cooperation.

Bicchieri's theory of social norms can offer a useful framework for understanding how descriptive norms of communication, coordination or cooperation emerge within collectives, including teams of AI or human-AI interactions. Descriptive norms can arise from individuals' observations of how others behave in specific situations, leading them to conform if they believe the behavior is widespread. In AI collectives or mixed human-AI teams, agents—whether human or artificial—may adjust their behavior based on what they perceive as common or acceptable practices. Descriptive norms might thus emerge from patterns of interaction that are repeated and become predictable, which leads to stable behavioral expectations.

For example, if a group of AI agents work together to optimize traffic flow in a smart city, over time, these AI agents may establish a pattern where they coordinate to reduce congestion by adjusting traffic signals in a synchronized manner. Each agent observes the others consistently contributing to this coordinated effort, which leads to smoother traffic flow. This coordinated behavior gradually becomes a descriptive norm among the AI agents: they "expect" that all agents will continue to coordinate in optimizing traffic, guiding their future interactions in similar scenarios.

Similarly, in a human-AI team setting, human operators and AI systems may work together to allocate medical resources in a hospital. Over time, a norm might emerge where human doctors consistently follow the AI's recommendations for resource allocation because the AI has shown high accuracy in predicting patient needs. This behavior is reinforced because the doctors observe that their colleagues also tend to trust the AI's decisions, and outcomes are generally positive. Over repeated interactions, the reliance on the AI's suggestions becomes a descriptive norm in the team.

In both cases, these norms—whether among AI agents or in mixed human-AI teams—create stable expectations about how decisions will be made. The AI agents in the traffic example expect their peers to continue coordinating their actions, and the doctors in the hospital example expect their colleagues to defer to the AI.

Such norms help streamline interactions but can also become resistant to change, even if circumstances shift, such as if the AI in the hospital begins making less accurate

European Innovation Council Funded by the European Union

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.

recommendations. As Bicchieri argues, shifts in social norms are often slow because they require a collective change in both what individuals believe others are doing (empirical expectations) and what they believe others think they should do (normative expectations). This is especially relevant in human-AI teams, where changing these expectations may necessitate reconfiguring how both humans and AI systems understand and respond to each other's roles and behaviors in the collective.

Table 4: overview of existing measures of collaborative awareness from social psychology.

Methods for assessing collaborative awareness from social psychology
Social psychologists use a combination of self-report measures, behavioral observations, and experimental paradigms to assess communication, coordination or cooperation, and the specific measures vary with the aspect being studied. Measures of group dynamics and team processes can assess coordination through role clarity scales, task interdependence measures or team mental model similarity indices. Social norms are usually assessed through verbal reports of empirical expectations (perceptions of how others typically behave) and injunctive expectations (perceptions of what behaviors others approve/disapprove of), while social norm adherence scales measure conformity to perceived norms.

7. Collaborative spatial awareness for robot swarms

How could one go about implementing collaborative spatial awareness in a swarm of robots? Moreover, what might be the concrete difference between a robot swarm capable of communication, coordination or cooperation along this single dimension?

A single agent within a distributed swarm could be given spatial awareness locally. This agent would then be aware of its own location with respect to either its environment or other moving objects within it (i.e. the rest of the swarm). This need not yet entail recognising the other moving agents as members of its own 'group' - but it very well could. If this is the case for all agents within the swarm on a local level, then the swam could eventually develop a shared reference frame. At the level of mere reactions to local agents (i.e. reacting to proximity), no shared reference frame is created. Only when there is more sophisticated exchange of information and model, such as through Gaussian Belief Propagation, a true cooperative shared reference frame can emerge (Jones and Hauert, 2024).

Robots can measure the relative location of other robots within their local sensing range and can also measure their own movement (odometry). These measurements constitute observations that are used to construct a local factor graph (Dellaert and Kaess, 2017) constraining possible robot positions. Robots within local sensing range also exchange messages that implement Gaussian Belief Propagation on the factor graphs, tying them together into a distributed graph that converges on the most consistent reconciliation of all the recent measurements made collectively across the swarm, summarised as the shared Distributed Reference Frame; each robot knows where it is relative to the swarm as a whole.

Within an intralogistics scenario a robot swarm might have to locate boxes within a bounded area and deliver these to a designated drop-off zone. Some boxes might require more than one agent to lift, whereas others could be delivered by a single agent. Success in performing this task could be measured in terms of overall speed to deliver all the boxes within the arena. The spatial awareness each individual agent has can be of different degrees, some might know the exact coordinates of a box, whereas others know only the quadrant of the arena in which a box is located. Moreover, we can imagine a swarm that immediately recognise which

European Innovation Council Funded by the European Union

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.

boxes need others to lift, and which can be done individually, but one could also imagine a swarm that would need to acquire and communicate this information also. If we are considering a swarm with heterogenous agents, such targeted communication might be essential to improve efficiency.

It is in such a scenario that genuine **communication**, beyond mere broadcasting becomes essential. If only certain members of the swarm can assist in the task at hand, then locating them and communicating specifically to those agents becomes especially pertinent. The firefighting case as detailed in the introduction illustrate the need for this. If we have a swarm consisting of domain-specific specialised agents a shared reference frame could be the stepping of point to enable more targeted communication as well as localisation.

If we imagine a scenario where two specialised heterogenous swarms are to collaborate in a shared space. Say drones spotting and putting out fires from the air, and robots on the ground creating firebreaks and reinforcing firefighters. Both swarms have their own reference frame relative to their own location, however it would be beneficial (if not crucial) for the two swarm to collaborate efficiently if they could share this.

Here we can distinguish between an egocentric and allocentric spatial reference frame (Klatzky, 1998). An egocentric perspective is one that is self-centred. Objects and their locations are represented relative to the observer's own position or viewpoint. For example, "the bookshelf is to my left" is an example of egocentric location information. An allocentric perspective is one that is world-centred, it refers to an external perspective. This means that in an allocentric reference frame points are located within a framework external to the observer and their position (Klatzky, 1998). It is a 'map-like' view of the environment, navigating using landmarks or other recognisable features such as the north pole. To extend the previous example, "the bookshelf is on my west" is an allocentric location information.

Within a single swarm each agent has an egocentric perspective which using Gaussian Belief Propagation, if all goes well, is consistent across all members of the swarm. However, for each they still view it from where they are within said swarm. On the level of a single swarm this is sufficient. However, if the swarm needs to collaborate with another swarm of a different 'species' (i.e. brand), or communication and work with human beings – then such an egocentric perspective might cease to be enough.

At this point, several intriguing questions arise regarding the nature of swarms. If two distinct swarms meet and converge on a shared reference frame, do they lose their distinct identities as separate entities? How we approach this question influences our understanding of the shared map —whether it embodies merely a sophisticated egocentric, a **quasi-allocentric,** or a clear case of an allocentric perspective.

This question invites future exploration of an **equilibrium** in the context of swarm dynamics. The balance between maintaining distinct identities and achieving functional integration reflect an ongoing negotiation, which can be more or less successful depending on the various architectures of the swarms involved, between individuality and cooperation. The granularity at which the scenario is analysed shapes our interpretation of the outcome.

Two swarms both working within the same map, could engage in continuously updating that map. In the firefighting scenario that could mean both updating the map of the environment from the air and the ground, but also monitoring the spread of the fire in a manner that is sharable between both swarms and involved human beings. In this case **coordination** could easily take place. If the drones are putting out the fire from the air by dropping water on it, the robots on the grounds might be creating a line and keeping the fire from spreading to especially flammable or sensitive areas. They have distinct tasks potentially even on different sides of

European Innovation Council
Funded by the European Union

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.

the forest, but to maximise efficiency need to **coordinate** in their updating of the map, as well as stay out of each other's way. However, if they are both working in the exact same area of the forest then **cooperation** is key, where their activities need to be aligned in their shared goal of putting out the fire.

How might this be achieved between two swarms? One approach would be, for example, distributed SLAM (e.g Lajoie et al. 2020). But this would seem to lose the separateness between the swarms and introduce a large amount of redundant and unnecessary information exchange and computation. Consider the scenario above, each swarm cares about different environmental information and builds a different 'map', the ground robots need to know about obstacles, sensitive areas, and potential firebreak locations, the drones need to know more about long range fire sensing, locations of charging stations and atmospheric conditions. But they do both care about some features in the global map – the location of fires. At the minimum then, each swarm must be able to agree on the transformation between their respective reference frames in order to share just the salient parts of each map.

One could argue that as long as both swarms' egocentric maps converge in the right places (e.g. fire locations), they may function within an emergent quasi-allocentric perspective. Each swarm maintains an egocentric view, but the convergence of these views for shared purposes introduces a level of abstraction that borders on allocentric – whilst maintaining the separation between the two swarms.

For the two physically separate swarms in the scenario highlighted above to agree between themselves about their respective transformation, they cannot rely on the intra-agent short range communication. One approach is to provide at least some members of each swarm with the ability to perform mutual long range inter-swarm measurement and communication. This could be used to support the minimal level of GBP processing to derive the transformation between the two reference frames.

In a more descriptive sense, each swarm is cohesive, with relatively large amounts of shorter-range intra-swarm observation and communication. Inter-swarm communication is far sparser but over longer distances, just sufficient to support the requirements for sharing of salient map features. Because of the sparser requirements, only a fraction of each entire swarm would need the more expensive long-range sensing and communication hardware.

Clearly, all agents within each swarm are **communicating**, in order to implement GBP and the swarm goals. Additionally, inter-swarm communication through the sparse agents is necessary to enable **coordination** with respect to the salient global information of fire locations. We don't talk here about **cooperation** on e.g. planning over the shared goal of fighting the fire**,** which would perhaps operate at a higher level of abstraction, but both the communication and coordination detailed above are necessary preconditions.

Conclusion

Collaborative awareness, as explored in this report, is essential for creating AI systems that can work safely and effectively with humans in shared environments. By focusing on the unique capabilities that emerge at the group level, this framework provides a conceptual basis for collaboration in AI collectives. We examined methods for delineating communication, coordination, and cooperation, drawing from ethology, game theory, neuroscience, social psychology, and swarm robotics.

Our dimensional model of collaborative awareness offers a clear approach to designing systems that adapt and work together in complex environments, ensuring ethical and reliable

European Innovation Council Funded by the European Union

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.

interactions. It emphasises the need for effective and meaningful assessments of artificial systems, evaluating their awareness profiles and their success in collaborating with diverse agents. This aims to enable the implementation of awareness in heterogeneous collectives of artificial systems in an ethically tractable manner, suitable for deployment in shared spaces with human beings.

European Innovation Council
Funded by the European Union

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.

References

Adriaans, P. (2024). Information. In: Zalta, E.N., Nodelman, U. (eds.). *The Stanford Encyclopedia of Philosophy* (Summer 2024 Edition). https://plato.stanford.edu/archives/sum2024/entries/information/

Ardón, P., Pairet, È., Lohan, K. S., Ramamoorthy, S., & Petrick, R. P. A. (2020). Affordances in Robotic Tasks—A Survey (Version 1). *arXiv.* https://doi.org/10.48550/ARXIV.2004.07400

Axelrod, R., & Hamilton, W. D. (1981). The Evolution of Cooperation. Science, 211(4489), 1390–1396. https://doi.org/10.1126/science.7466396

Babb, S. J., & Crystal, J. D. (2006). Episodic-like Memory in the Rat. *Current Biology*, 16(13), 1317–1321. https://doi.org/10.1016/j.cub.2006.05.025

Bardsley, N., Mehta, J., Starmer, C., and Sugden, R. (2010). Explaining focal points: Cognitive hierarchy theory versus team reasoning. The Economic Journal 120, 40–79.

Bacharach, M. (2006). *Beyond individual choice: teams and frames in game theory*. Princeton University Press.

Bayne, T. (2011). Agentive Experiences as Pushmi-Pullyu Representations. In: Aguilar, J.H., Buckareff, A.A., Frankish, K. (eds) *New Waves in Philosophy of Action.* New Waves in Philosophy. Palgrave Macmillan, London. https://doi.org/10.1057/9780230304253_11

Bayne, T., Hohwy, J., & Owen, A. M. (2016). Are There Levels of Consciousness? *Trends in Cognitive Sciences,* 20(6), 405–413. https://doi.org/10.1016/j.tics.2016.03.009

Bayne, T., & Pacherie, E. (2007). Narrators and comparators: The architecture of agentive self-awareness. *Synthese,* 159(3), 475–491. https://doi.org/10.1007/s11229-007-9239-9

Bshary, R., & Raihani, N. J. (2017). Helping in humans and other animals: A fruitful interdisciplinary dialogue. *Proceedings of the Royal Society B: Biological Sciences,* 284(1863), 20170929. https://doi.org/10.1098/rspb.2017.0929

Bicchieri, C. (2006). *The grammar of society.* Cambridge University Press.

Bicchieri, C. (2017). *Norms in the Wild: How to Diagnose, Measure, and Change Social Norms.* Oxford University Press. https://doi.org/10.1093/acprof:oso/9780190622046.001.0001

Binmore, K. (2005). *Natural justice.* Oxford University Press.

Birch, J., Schnell, A. K., & Clayton, N. S. (2020). Dimensions of Animal Consciousness. *Trends in Cognitive Sciences,* 24(10), 789–801. https://doi.org/10.1016/j.tics.2020.07.007

Birk, A., & Carpin, S. (2006). Merging occupancy grid maps from multiple robots. *Proceedings of the IEEE: Special Issue on Multi-Robot Systems*, 94(7), 1384–1387.

Bones, H., Ford, S., Hendery, R., Richards, K., & Swist, T. (2021). In the Frame: The Language of AI. *Philosophy & Technology*, 34(S1), 23–44. https://doi.org/10.1007/s13347-020-00422-7

Browning, H. (2023). Welfare comparisons within and across species. *Philosophical Studies*, 180(2), 529-551.

European Innovation Council Funded by the European Union

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.

Buckner, C. (2014). The Semantic Problem(s) with Research on Animal Mind-Reading. *Mind & Language*, 29(5), 566–589. https://doi.org/10.1111/mila.12066

Byrne, R. W., Cartmill, E., Genty, E., Graham, K. E., Hobaiter, C., & Tanner, J. (2017). Great ape gestures: Intentional communication with a rich set of innate signals. *Animal Cognition*, 20(4), 755–769. https://doi.org/10.1007/s10071-017-1096-4

Call, J. (2009). Contrasting the Social Cognition of Humans and Nonhuman Apes: The Shared Intentionality Hypothesis. *Topics in Cognitive Science,* 1(2), 368–379. https://doi.org/10.1111/j.1756-8765.2009.01025.x

Camerer, C. (2003). *Behavioral game theory: Experiments in strategic interaction.* Princeton University Press.

Chemero, A., & Turvey, M. T. (2007). Gibsonian Affordances for Roboticists. *Adaptive Behavior,* 15(4), 473–480. https://doi.org/10.1177/1059712307085098

Chiba, A. A., & Krichmar, J. L. (2020). Neurobiologically Inspired Self-Monitoring Systems. Proceedings of the IEEE, 108(7), 976–986. https://doi.org/10.1109/JPROC.2020.2979233

Clark, A. (2015). Embodied Prediction. *Open MIND.* https://doi.org/10.15502/9783958570115

Clayton, P., & Davies, P. C. W. (Eds.). (2006). *The re-emergence of emergence: The emergentist hypothesis from science to religion.* Oxford University Press.

Clayton, N. S., & Dickinson, A. (1998). Episodic-like memory during cache recovery by scrub jays. *Nature*, 395(6699), 272-274.

Clutton-Brock, T. (2009). Cooperation between non-kin in animal societies. *Nature*, 462(7269), 51–57. https://doi.org/10.1038/nature08366

Cobianchi, L., Gigliuto, C., De Gregori, M., Malafoglia, V., Raffaeli, W., Compagnone, C., Visai, L., Petrini, P., Avanzini, M. A., Muscoli, C., Calabrese, F., Dominioni, T., Allegri, M., & Vigano, J. (2014). Pain assessment in animal models: Do we need further studies? *Journal of Pain Research,* 227. https://doi.org/10.2147/JPR.S59161

Crockford, C., Wittig, R. M., Mundry, R., & Zuberbühler, K. (2012). Wild Chimpanzees Inform Ignorant Group Members of Danger. *Current Biology*, 22(2), 142–146. https://doi.org/10.1016/j.cub.2011.11.053

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

Davies, J. R., Garcia-Pelegrin, E., Baciadonna, L., Pilenga, C., Favaro, L., & Clayton, N. S. (2022). Episodic-like memory in common bottlenose dolphins. *Current Biology*, 32(15), 3436-3442.e2. https://doi.org/10.1016/j.cub.2022.06.032

Dellaert, F. and Kaess, M. (2017), Factor Graphs for Robot Perception, *Foundations and Trends® in Robotics*: Vol. 6: No. 1-2, pp 1-139. http://dx.doi.org/10.1561/2300000043

Deroy, O. (2023). The Ethics of Terminology: Can We Use Human Terms to Describe AI? *Topoi,* 42(3), 881–889. https://doi.org/10.1007/s11245-023-09934-1

Deroy, O., Bacciu, D., Bahrami, B., Della Santina, C., & Hauert, S. (2024). Shared Awareness Across Domain-Specific Artificial Intelligence: An Alternative to Domain-General

European Innovation Council Funded by the European Union

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.

Intelligence and Artificial Consciousness. *Advanced Intelligent Systems*, 2300740. https://doi.org/10.1002/aisy.202300740

Deshpande, A., Van De Waal, E., & Zuberbühler, K. (2023). Context-dependent alarm responses in wild vervet monkeys. *Animal Cognition*, 26(4), 1199–1208. https://doi.org/10.1007/s10071-023-01767-0

Dorsch, J., & Deroy, O. (2024a). The impact of labeling automotive AI as 'trustworthy' or 'reliable' on user evaluation and technology acceptance (Version 1). *arXiv.* https://doi.org/10.48550/ARXIV.2408.10905

Dorsch, J., & Deroy, O. (2024b). "Quasi-Metacognitive Machines: Why We Don't Need Morally Trustworthy AI and Communicating Reliability is Enough." *Philosophy and Technology*, 37(2). https://doi.org/10.1007/s13347-024-00752-w

Drugowitsch, J., Mendonça, A. G., Mainen, Z. F., & Pouget, A. (2019). Learning optimal decisions with confidence. *Proceedings of the National Academy of Sciences*, 116(49), 24872–24880.

Dung, L., & Newen, A. (2023). Profiles of animal consciousness: A species-sensitive, two-tier account to quality and distribution. *Cognition,* 235, 105409. https://doi.org/10.1016/j.cognition.2023.105409

Falk, A., Fehr, E., and Fischbacher, U. (2003). On the nature of fair behavior. *Economic Inquiry* 41, 20–26.

von Frisch, K. (1954). *The Dancing Bees: An Account of the Life and Senses of the Honey Bee.* Springer Vienna. https://doi.org/10.1007/978-3-7091-4697-2

Floridi, L. (2011). *The Philosophy of Information.* Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199232383.001.0001

Frith, C. D. (2012). The role of metacognition in human social interactions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1599), 2213–2223. https://doi.org/10.1098/rstb.2012.0123

Gallistel, C. R., & Gibbon, J. (2000). Time, rate, and conditioning. *Psychological Review,* 107(2), 289–344. https://doi.org/10.1037/0033-295X.107.2.289

Gibson, J. J. (1979). *The ecological approach to visual perception*. Houghton, Mifflin and Company.

Hamilton, W. D. (1964). The genetical evolution of social behaviour. II. *Journal of Theoretical Biology*, 7(1), 17–52. https://doi.org/10.1016/0022-5193(64)90039-6

Harsanyi, J. C. and Selten, R. (1988). *A general theory of equilibrium selection in games.* MIT Press.

Hauert, S., Mitri, S., Keller, L., & Floreano, D. (2014). Evolving Cooperation: From Biology to Engineering. In *The Horizons of Evolutionary Robotics* (pp. 203). Massachusetts Institute of Technology (MIT) Press.

Heyes, C. Animal mindreading: what's the problem?. *Psychon Bull Rev* 22, 313–327 (2015). https://doi.org/10.3758/s13423-014-0704-4

Horton, T. E., Chakraborty, A., & Amant, R. S. (2012). Affordances for robots: a brief survey. Avant: Trends in Interdisciplinary Studies 3 (2):70-84.

European Innovation Council Funded by the European Union

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.

Innocente, M. S., & Grasso, P. (2019). Self-organising swarms of firefighting drones: Harnessing the power of collective intelligence in decentralised multi-robot systems. *Journal of Computational Science*, 34, 80–101. https://doi.org/10.1016/j.jocs.2019.04.009

Isoni, A., Poulsen, A., Sugden, R., and Tsutsui, K. (2013). Focal points in tacit bargaining problems: Experimental evidence. *European Economic Review* 59, 167–188.

Isoni, A., Poulsen, A., Sugden, R., and Tsuitsui, K. (2019) Focal points and payoff information in tacit bargaining. *Games and Economic Behavior* 114, 193–214.

Johnson, B. (2022). Metacognition for artificial intelligence system safety – An approach to safe and desired behavior. *Safety Science*, 151, 105743. https://doi.org/10.1016/j.ssci.2022.105743

Jones, S., & Hauert, S. (2023). Frappe: Fast fiducial detection on low cost hardware. *Journal of Real-Time Image Processing*, 20(6), 119. https://doi.org/10.1007/s11554-023-01373-w

Jones, S., & Hauert, S. (Accepted/In press). Distributed Spatial Awareness for Robot Swarms. Paper presented at The International Symposium on Distributed Autonomous Robotic Systems (DARS) 2024, New York City, New York, United States.

Kalkman, D. (2019). New problems for defining animal communication in informational terms. *Synthese*, 196(8), 3319–3336. https://doi.org/10.1007/s11229-017-1598-2

Karpus, J. and Radzvilas, M. (2018) Team reasoning and a measure of mutual advantage in games. *Economics and Philosophy* 34, 1–30.

Klatzky, R.L. (1998). Allocentric and Egocentric Spatial Representations: Definitions, Distinctions, and Interconnections. In: Freksa, C., Habel, C., Wender, K.F. (eds) *Spatial Cognition.* Lecture Notes in Computer Science(), vol 1404. Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-69342-4_1

Lajoie, P. -Y., Ramtoula, B., Chang, Y., Carlone, L. and Beltrame, G. DOOR-SLAM: Distributed, Online, and Outlier Resilient SLAM for Robotic Teams *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 1656-1663, April 2020, doi: 10.1109/LRA.2020.2967681

Le Poidevin, R. (2019). The Experience and Perception of Time. *The Stanford Encyclopedia of Philosophy* (Summer 2019 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/sum2019/entries/time-experience/>.

Lewes, G. H. (1877). *Problems of life and mind* (Vol. 2). Trübner & Company.

Lourenco, I., Ventura, R., & Wahlberg, B. (2020). Teaching Robots to Perceive Time: A Twofold Learning Approach. *2020 Joint IEEE 10th International Conference on Development and Learning and Epigenetic Robotics* (ICDL-EpiRob), 1–7. https://doi.org/10.1109/ICDL-EpiRob48136.2020.9278033

Lurz, R. W. (2011). *Mindreading animals: The debate over what animals know about other minds*. MIT Press.

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

Maniadakis, M., Aksoy, E. E., Asfour, T., & Trahanias, P. (2016). Collaboration of heterogeneous agents in time constrained tasks. *2016 IEEE-RAS 16th International*

European Innovation Council Funded by the European Union

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.

Conference on Humanoid Robots (Humanoids), 448–453. https://doi.org/10.1109/HUMANOIDS.2016.7803314

Maniadakis, M., Hourdakis, E., Sigalas, M., Piperakis, S., Koskinopoulou, M., & Trahanias, P. (2020). Time-Aware Multi-Agent Symbiosis. *Frontiers in Robotics and AI,* 7, 503452. https://doi.org/10.3389/frobt.2020.503452

Maynard Smith, J. (1964). Group selection and kin selection. *Nature* 201, 1145–1147.

Maynard Smith, J. (1982). *Evolution and the theory of games.* Cambridge University Press.

Maynard Smith, J., & Price, G. R. (1973). The Logic of Animal Conflict. *Nature*, 246(5427), 15–18. https://doi.org/10.1038/246015a0

McConville, A., Tzoumas, G., Salinas, L. R., Munera, M., & Hauert, S. (2024). Adoption of UAV Swarm Technology: Survey and Opinions of Firefighters. *2024 IEEE International Conference on Advanced Robotics and Its Social Impacts* (ARSO), 228–234. https://doi.org/10.1109/ARSO60199.2024.10557806

McMillan, C. T., Rascovsky, K., Khella, M. C., Clark, R., & Grossman, M. (2012). *The neural basis for establishing a focal point in pure coordination games*. Social Cognitive and Affective Neuroscience, 7(8), 881–887. https://doi.org/10.1093/scan/nsr070

Mehta, J., Starmer, C., and Sugden, R. (1994). The nature of salience: An experimental investigation of pure coordination games. *The American Economic Review* 84, 658–673.

Melis, A. P., & Semmann, D. (2010). How is human cooperation different? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1553), 2663–2674. https://doi.org/10.1098/rstb.2010.0157

Morwald, T., Zillich, M., Prankl, J., & Vincze, M. (2011). Self-monitoring to improve robustness of 3D object tracking for robotics. 2011 *IEEE International Conference on Robotics and Biomimetics*, 2830–2837.

Mylopoulos, M. (2017). A cognitive account of agentive awareness. *Mind & Language*, 32(5), 545–563. https://doi.org/10.1111/mila.12158

Newcombe, N. S., Balcomb, F., Ferrara, K., Hansen, M., & Koski, J. (2014). Two rooms, two representations? Episodic-like memory in toddlers and preschoolers. *Developmental Science*, 17(5), 743–756. https://doi.org/10.1111/desc.12162

Proust, J. (2014). Metacognition and mindreading: one or two functions. In: Beran, M. J., Brandl, J., Perner, J., & Proust, J. *Foundations of Metacognition* (234-251). OUP Oxford.

Radzvilas, M. and Karpus, J. (2021). Team reasoning without a hive mind. *Research in Economics* 75, 345–353.

Roldán-Gómez, J. J., González-Gironda, E., & Barrientos, A. (2021). A Survey on Robotic Technologies for Forest Firefighting: Applying Drone Swarms to Improve Firefighters' Efficiency and Safety. *Applied Sciences*, *11*(1), 363. https://doi.org/10.3390/app11010363

Saeedi, S., Trentini, M., Seto, M., & Li, H. (2016). Multiple-Robot Simultaneous Localization and Mapping: A Review: Multiple-Robot Simultaneous Localization and Mapping. *Journal of Field Robotics*, 33(1), 3–46. https://doi.org/10.1002/rob.21620

European Innovation Council Funded by the European Union

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.

Schel, A. M., Townsend, S. W., Machanda, Z., Zuberbühler, K., & Slocombe, K. E. (2013). Chimpanzee Alarm Call Production Meets Key Criteria for Intentionality. *PLoS ONE*, 8(10), e76674. https://doi.org/10.1371/journal.pone.0076674

Schelling, T. C. (1960). *The Strategy of Conflict.* Cambridge and London: Harvard University Press.

Seraj, E., Silva, A., & Gombolay, M. (2019). Safe Coordination of Human-Robot Firefighting Teams (arXiv:1903.06847). *arXiv.* http://arxiv.org/abs/1903.06847

Seyfarth, R. M., Cheney, D. L., & Marler, P. (1980). Vervet monkey alarm calls: Semantic communication in a free-ranging primate. *Animal Behaviour,* 28(4), 1070–1094. https://doi.org/10.1016/s0003-3472(80)80097-2

Seyfarth, R. M., & Cheney, D. L. (2003). Signalers and Receivers in Animal Communication. *Annual Review of Psychology*, 54(1), 145–173.

Skyrms, B. (2004). *The Stag Hunt and the evolution of social structure.* Cambridge University Press.

Skyrms, B. (2010). *Signals: Evolution, learning, & information.* Oxford University Press.

Soares, S., Atallah, B. V., & Paton, J. J. (2016). Midbrain dopamine neurons control judgment of time. *Science*, 354(6317), 1273–1277. https://doi.org/10.1126/science.aah5234

Soter, G., Conn, A., Hauser, H., & Rossiter, J. (2018). Bodily Aware Soft Robots: Integration of Proprioceptive and Exteroceptive Sensors. *2018 IEEE International Conference on Robotics and Automation* (ICRA), 2448–2453. https://doi.org/10.1109/ICRA.2018.8463169

Spencer, R. W. (2024). The Building Blocks of Consciousness (Version 1). *arXiv*. https://doi.org/10.48550/ARXIV.2405.06075

Sperber, D., & Mercier, H. (2018). Why a modular approach to reason? *Mind & Language*, 33(5), 533–541. https://doi.org/10.1111/mila.12208

Stachowicz, D., & Kruijff, G.-J. M. (2012). Episodic-Like Memory for Cognitive Robots. *IEEE Transactions on Autonomous Mental Development,* 4(1), 1–16. https://doi.org/10.1109/TAMD.2011.2159004

Strasser, A. (2018). Minimal Mindreading and Animal Cognition. *Grazer Philosophische Studien*, 95(4), 541-565. https://doi.org/10.1163/18756735-000048

Sugden, R. (1993). Thinking as a team: towards an explanation of nonselfish behavior. *Social Philosophy and Policy* 10, 69–89.

Tomasello, M., & Carpenter, M. (2007). Shared intentionality. *Developmental Science,* 10(1), 121–125. https://doi.org/10.1111/j.1467-7687.2007.00573.x

Tomasello, M. (2022). The coordination of attention and action in great apes and humans. *Philosophical Transactions of the Royal Society B: Biological Sciences,* 377(1859). https://doi.org/10.1098/rstb.2021.0093

Tzoumas, G., Pitonakova, L., Salinas, L., Scales, C., Richardson, T., & Hauert, S. (2023). Wildfire detection in large-scale environments using force-based control for swarms of UAVs. *Swarm Intelligence,* 17(1–2), 89–115. https://doi.org/10.1007/s11721-022-00218-9

Tzoumas, G., Salinas, L., McConville, A., Richardson, T., & Hauert, S. (2024). Use case design for swarms of firefighting UAVs via mutual shaping. *2024 IEEE International*

Funded by the European Union

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.

Conference on Advanced Robotics and Its Social Impacts (ARSO), 43–48. https://doi.org/10.1109/ARSO60199.2024.10558013

Varela, F. J. (1999). Present-time consciousness. *Journal of consciousness studies,* 6(2-3), 111-140.

Visco-Comandini, F., Ferrari-Toniolo, S., Satta, E., Papazachariadis, O., Gupta, R., Nalbant, L. E., & Battaglia-Mayer, A. (2015). Do non-human primates cooperate? Evidences of motor coordination during a joint action task in macaque monkeys. *Cortex,* 70, 115–127. https://doi.org/10.1016/j.cortex.2015.02.006

Voinov, P. V., Call, J., Knoblich, G., Oshkina, M., & Allritz, M. (2020). Chimpanzee Coordination and Potential Communication in a Two-touchscreen Turn-taking Game. *Scientific Reports*, 10(1). https://doi.org/10.1038/s41598-020-60307-9

Wilson, E. O. (2000). *Sociobiology: The new synthesis (25th anniversary ed*). Belknap Press of Harvard University Press.

Winfield, A. (2019). Ethical standards in robotics and AI. *Nature Electronics*, 2(2), 46–48. https://doi.org/10.1038/s41928-019-0213-6

European Innovation Council
Funded by the European Union

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.

Appendix A: Glossary

Awareness: The situatedness of an artificial system in a multidimensional space of action-perception abilities (Meertens, under review).

Awareness refers to the capacities these systems can be given to enable adaptation and navigation in dynamic, complex environments. It also encapsulates the system's or collective's degree of interaction with itself, others, and its environment, structured along various dimensions.

Degree of awareness: awareness is approached as a graded phenomenon, rather than an on/off property, or a hierarchical levelled approach.

Awareness profile: A unique configuration representing an artificial system's awareness across multiple dimensions, offering a snapshot of its action-perception abilities at a given time.

Dimensions of awareness: Awareness does not vary along a singular scale but across multiple, distinct aspects or "dimensions." These dimensions allow comparison between different systems, reflecting varying abilities like spatial, temporal, or meta-cognitive awareness. The use of dimensions permits flexible, non-linear comparisons, avoiding assumptions of a universal standard for awareness.

Capacity or **ability:** A multi-track dispositional property indicating the potential for success in certain tasks. Each dimension of awareness is tied to a set of abilities that a system or agent possesses, which enables the successful execution of tasks associated with that ability.

Collaborative awareness: the time-bound adaptive pursuit of multi-agent goals, including through changes in the environment and group rates or compositions.

Levels of collaboration: Collaboration encompasses a range of behaviors, from simple communication to more complex coordination and cooperation. These levels, while not strictly hierarchical, are often interdependent, with more sophisticated levels building upon foundational ones.

Communication: The capacity for sharing information between agents, ranging from simple signals or broadcasts to more complex, interactive exchanges. Communication can occur at varying levels of sophistication, depending on the agents' abilities and goals.

Coordination: occurs when agents with distinct, yet interdependent, goals work together to ensure that their activities align for mutual benefit.

Cooperation: occurs when agents share a common goal, leading to the alignment of their activities to achieve that shared objective.

Agency: The system's ability to exert control over its actions and outcomes, often tied to counterfactual thinking (the capacity to act otherwise) and causal influence on the environment or other agents.

Reliability: The consistency with which a system or agent performs its tasks or meets expectations. Unlike trustworthiness, which implies normative reasoning, reliability is based on observable and measurable performance over time.

Confidence: The system's self-assessment of its ability to succeed or achieve a goal. Confidence involves evaluating the certainty or uncertainty related to specific actions or decisions.

European Innovation Council Funded by the European Union

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.

Explainability: The degree to which a system can make its processes, decisions, and actions understandable to human users. It emphasizes transparency and clarity, ensuring that the system's behaviour can be interpreted and followed.

Justifiability: The system’s ability to provide reasons or rationale for its decisions and actions, particularly when questioned. Justifiability connects explainability to a normative framework, ensuring actions are not only understood but ethically or logically sound.

European Innovation Council
Funded by the European Union

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.