Philosophical Problems in Science

No 79
2025

Philosophical Problems in Science

Zagadnienia Filozoficzne w Nauce

e-ISSN 2451-0602 (electronic format)
ISSN 0867-8286 (print format, discontinued)

Editorial Office

Philosophical Problems in Science
(Zagadnienia Filozoficzne w Nauce)
e-mail: info@zfn.edu.pl
www.zfn.edu.pl

Publisher: Copernicus Center Foundation
Pl. Szczepański 8, 31-011 Kraków POLAND
tel. (+48) 12 422 44 59
e-mail: info@copernicuscenter.edu.pl
www.copernicuscenter.edu.pl

Philosophical Problems in Science

Zagadnienia Filozoficzne w Nauce

79 — 2025

Articles

Marzena Bielecka, Andrzej Bielecki
The problem of scene understanding by artificial autonomous agents 3

Floriana Ferro, Alessandro Giuseppe Privitera, Andrea Gulli, Michele Geronazzo
From Digital Twins to Entangled Twins: A phenomenological model for Industry 5.0 21

Roman Krzanowski
Consciousness, normativity, and intrinsic meaning in large language models 41

Teresa Marcinów
Towards happiness: Ancient Greek values in the age of computational intelligence supporting happiness—selected issues 49

Anna Sarosiek
From simple rules to quasi intentionality: emergent coordination and collective cognition 59

Miroslav Vacura
More than a tool: Artificial Intelligence in a philosophical history of technologies 73

Articles

Philosophical Problems in Science (2025), **79**, 3–19
(Zagadnienia Filozoficzne w Nauce)
doi:10.59203/zfn.79.732

ARTICLE

The problem of scene understanding by artificial autonomous agents

Marzena Bielecka* and Andrzej Bielecki

AGH University of Krakow
*Corresponding author. Email: bielecka@agh.edu.pl

Abstract

In this paper, the problem of scene understanding in the context of autonomous robotics is considered. The implemented solutions, based on various types of artificial intelligence systems, are described and the perspectives of their development are discussed. In particular, the problem of cognition that reflects the structure and nature of the environment in which the artificial autonomous agent operates, is studied. The specifics of scene models and algorithms for their construction and verification of their adequacy are discussed. The presentation of the existing partial solutions, including the results obtained by the authors, is the starting point of the analysis. Possibilities of new approaches are analysed, especially those that are biologically inspired. The potential fundamental limitations are studied. These problems are discussed in the context of epistemology and in the frame of cybernetics. The analysis of the relationship between philosophy and artificial intelligence in the aspect of cognitive vision systems of autonomous robots is the aim of this paper.

Keywords: epistemology, autonomous agent, scene understanding, cognition, robotic autonomy

1. Introduction

Until recently, the human being was the only cognitive subject analysed on philosophical grounds in the context of epistemology. However, the rapid development of artificial intelligence implies, among others, its wide application in embodied systems such as artificial autonomous agents, for instance, autonomous robots and autonomous unmanned aerial vehicles (Bielecki 2024b; Floreano and Wood 2015). Currently, artificial autonomous agents are characterized by only limited autonomy. A prerequisite for achieving full autonomy is adequate modelling of cognitive functions that are associated with sensory, creating, for example, the cognitive vision system (Tadeusiewicz 1992). The full understanding of the environment in which the agent operates is one of the aspects of the said functions. Such understanding consists of recognizing individual objects, their mutual spatial relationships, linking them with the appropriate semantics and cognitive models the agent is equipped with, for instance, by comparing the perceived scene with an incorporated terrain map. It also includes the integration of signals from various types of sensors and the creation of

environment models that will be a reference point in the process of understanding the scene (Katsev et al. 2011; Li et al. 2017; Wang et al. 2017). The experience gained so far allows us to assume that drawing ideas from biological prototypes may be an effective approach to developing systems for understanding the environment (Hu et al. 2017; Siagian and Itti 2009; Tejera et al. 2018) which, among others, can be a starting point for human-robot interactions (Kang and Han 2023; Quintas 2023; Umbrico et al. 2023).

The development, implementation, and verification of efficiency of robotic cognitive systems can shed new light on classic philosophical issues. These include the concept of truth and its verification, the reference of the cognitive subject to physical reality, information processing within the system in the context of the nature of information itself, and testing the adequacy of the model from an ontological and functionalist perspective. These problems are situated in an intensively explored stream of studies that concerns the relationship between philosophy and technology (Polak 2023).

While contemporary advancements in artificial intelligence have led to the emergence of sophisticated architectures, such as vision—language models (VLM) and affordance—learning agents, these systems do not necessarily circumvent the fundamental epistemic barriers discussed in this work. Modern VLMs, for instance, despite their high performance in object categorization, primarily operate on statistical correlations between visual and linguistic patterns. From a Kantian perspective, as analysed in this paper, they remain within the realm of establishing relationships between phenomena without achieving the capacity for autonomous conceptualization—the ability to create the idea behind the category. Furthermore, although affordance-learning agents represent a significant step toward functional scene analysis, they often lack a formal, transparent mechanism for verifying the adequacy of their internal models. The cybernetic approach presented here suggests that genuine understanding requires the agent to not only detect a functional property, but also to verify it through a goal-oriented process where predicted internal states are compared with achieved states. Without this self-reflective loop and the capacity for conceptualization, even the most advanced contemporary systems remain at the level of complex signal processing and pattern recognition rather than achieving the higher-order scene understanding characteristic of cognitive subjects.

The paper is organised in the following way. In the next section, the problem of robotic cognition is discussed using robotic cognitive vision systems as an example. In particular, the levels of cognition are specified in the frame of cybernetics. Furthermore, the open problems are presented. In Section 3, the current status of the scene understanding problem is discussed on the basis of selected examples of vision cognitive systems of the robots as well as the problem of learning. Philosophical aspects of the problem are analysed in Section 4.

2. The formulation of the problem

An autonomous cybernetic agent operating in a given environment has to be equipped with the system of sensors. The signals detected by sensors are then processed and analyzed at various levels. From a cybernetic point of view, the following levels in this analysis should be specified:

1. **Signal processing**. Signals received from the agent's sensor network, especially when they are complex in nature, such as camera images, must be pre-processed in order to

enhance or extract their essential features. Denoising, contrasting, binarization, skeletonization can be given as examples of such operations.

2. **Signal analysis** is the lower level of sensor data analysis. It allows the agent, for example, to quickly detect changes in a data stream. The detection of motion on the image from a camera monitoring the surroundings of the house can be put as an example.
3. **Simple classification.** At this level, the object is classified into one of the predefined classes on the basis of a feature that can be verified directly. Classification of a fruit as ripe or unripe based on its color is an example of such classification (Kurpaska et al. 2021).
4. **Pattern recognition.** At this level, a structure of single objects is analyzed. First of all, a representation of the pattern should be defined, usually formally. Then, the features that characterize the object are required. Usually, the preselected features are transformed into the more optimal form, for instance by reducing their dimension. The selection of the features, that have the most discriminative power and are sufficient to describe the recognized objects unequivocally, is the last stage of the recognition system design—see (Flasiński 2016, Section 10.1). In complex objects, individual features often have a structural character and in order to specify them for a given object, a structural analysis of the object should be performed, for instance an analysis of the contour of its image. Such complex analysis is often done by using syntactic methods—see (Flasiński 2019, Section 1.3). In order to analyze a given object, the preprocessing of the image of the object must be done. It consists in, for instance, normalization and scaling. Usually, classification of an object is the result of pattern recognition.
5. **Scene analysis** includes the identification of objects and the relationships between them. In the case of complex scenes, it covers complex issues that require multi-level analysis. An example is the problem of deciding whether a given line is the edge of an object or the edge of its shadow. For this purpose, it is necessary to analyze the positions of light sources and obstacles that reflect, absorb or scatter light. Scene analysis is a necessary step to be able to move on to the level of understanding the scene.
6. **Scene understanding** consists in associating its structure with certain meanings. They can be both structural and functional. For example, detecting the presence of other agents may mean the possibility of establishing contact with them, forming a coalition in order to perform the task collectively, but it may also mean a threat if the analyzed scene is a war zone.

One subtlety should be noticed. Namely, the division into objects (pattern recognition level) and the scene is common in computer science in the context of artificial intelligence. From a philosophical point of view, it is ontological in nature and is, in a way, arbitrary. The related philosophical problems are analysed in (Krzanowski and Polak 2022).

At the current stage of development of computer science, in particular artificial intelligence systems, the algorithms implementing points 1–3 are well developed, and the algorithms implementing points 4–5 are able to solve a wide class of tasks quite effectively. However, the algorithms implementing point 6 are in the very early stages of development. Accordingly, the following questions arise:

1. What are the perspectives for the development of scene understanding systems?
2. To what extent can systems of this type be based on the modelling of structures and processes in biological agents? In particular, how helpful in the development of scene

understanding algorithms can be the achievements of psychology? These two questions are legitimate and interesting from the scientific point of view in the light of the fact that the problem of understanding the scene is closely related to the autonomy of the agent, and this, in turn, is discussed in the context of biology. It should be noted that an in-depth analysis of the autonomy of biological agents is available (Rosslenbroich 2014) as well as the problem of biological basis of the autonomy, first of all development of neural systems and, as a consequence, sensorics and information processing (Scott 1995; Tadeusiewicz 2010).

3. What are the possible limitations in artificial systems of this type?

Discussing these issues in the context of artificial autonomous agent vision systems, is presented in the next section. Analysis is based on exemplary vision systems of robots and is conducted in the context of epistemology.

3. Cognitive scene understanding by artificial agents—the current status and immediate prospects

In this section, the problem of cognition that reflects the structure and nature of the environment in which the artificial autonomous agent operates is studied. In general, the philosophical systems in which truth as such is analyzed can be divided into *closed*, such as mathematics, existing *for itself* and open, i.e., those that operate in a certain environment, which is their *world*. In open systems, truth is related to learning about the world. Creating the models of the world is an effect of this learning. Cybernetics provides tools that shed significant and new light on the procedures of cognition by autonomous (in the cybernetic sense) agents, creation of models of the world, and verification of these models by autonomous agents that are the cognitive subjects.

In this section, we discuss the specifics of scene models and algorithms for their construction and verification of their adequacy. The presentation of the exemplary existing partial solutions is the starting point of the analysis. The possibilities of their development, generalization and integration are then discussed.

3.1 Cognition

The creation of an adequate model of the analyzed scene or, more generally, a model of the environment in which the agent operates is one of the crucial steps in the problem of the scene understanding. If the agent is fully autonomous, it must create the said models on its own. The models can be created on the basis of the signals obtained from the environment as well as on the cognitive frames defined in the agent's cognitive module. The model creation includes the following steps.

1. **Creating models of individual objects** by using the information created on the basis of signals received from the agent's sensory network. An important and often overlooked problem is the generation of information based on the acquired signals. A significant problem is the lack of an adequate theory of information. Existing theories, known as information theories (for instance, Shanon information theory, Kolmogorov information theory), are generally concerned with a method of measuring or calculating the amount of information rather than its generation or its nature, which is essential for future

development of algorithms for generating information from signals. However, it should be noted that contemporary frameworks, such as active inference and information-theoretic active perception, provide a bridge between signal processing and semantic understanding. In these approaches, information is not merely a passive input but a result of the agent's actions aimed at minimizing uncertainty (entropy) regarding its environment (Bajcsy, Aloimonos, and Tsotsos 2018; Friston 2009). Consequently, the algorithm for active investigation of the object by an agent can be viewed as a mechanism for maximizing information gain. This process allows the agent to update its internal cognitive frames and verify the adequacy of its scene models through purposeful interaction, effectively linking action with semantic exploration. The concept of structural information (Bielecka et al. 2025; Bielecki and Schmittel 2022; Bielecki and Stocki 2023), as well as the currently discussed theories of physical information (Barreiro et al. 2020; Krzanowski 2020, 2022; Roterman and Konieczny 2023), are a good starting point for the analysis of the problem of information generation. Such an algorithm should be based on active investigation of the object by an agent by using its sensors. After integrating the signals from various sensors, the model would be built. Such an approach was tested both for the two-dimensional and three-dimensional models in the context of unmanned aerial vehicles and primary results are promising (Bielecki et al. 2016; Bielecki, Buratowski, and Śmigielski 2013, 2014). For instance, let us briefly recall the algorithm for creating a 3-D representation of a single building by an unmanned aerial agent (Bielecki, Buratowski, and Śmigielski 2014). It is assumed that the agent operates in an urban-type area. The general idea is to construct the set of walls that constitutes 3-dimensional model of the analysed object. To enrich the model and make it more complete, the algorithm also creates representation of holes in the investigated building. To create 3-dimensional representation, the proposed algorithm is based on the idea of composing the solid from projections that consists of photographed and vectorized sides of a building. One photo is taken from above in order to show the top side of the investigated building. Next, the building is photographed horizontally towards half of the sides represented by the convex hull of the top side shape. This allows the agent to detect holes that are regarded as the object features. Such an algorithm allows the agent to create the solid representation. The mentioned integration may include, in addition to the image from the camera, also e.g., lidar signals.

2. **Modeling the spatial relationships between individual objects** is a key element of scene representation, and thus its analysis and, as a consequence, understanding. The said spatial relations are usually modelled by using graph-based models (Flasiński 2019), section 12.1, sometimes aided with algebraic methods (Chaib-draa 2002). Let us consider the case where the agent flies at high altitude—see Bielecki and Śmigielski (2017) for details. Since the artificial flying agent operates at high altitudes, from its point of view, the studied scene is two-dimensional. It is assumed that the robot operates in the urban-type environment. The scene is represented as a neighborhood graph that encodes data about the buildings shapes, locations, and spatial relations. The robot understands parts of the examined scene by analyzing the context of the objects' neighborhoods. Such a scene representation can be used effectively for recognition of the object, while a few objects of the same shape exist in the scene because during the recognition process of an individual building, not only the information about the shape of the object is exploited, but also the spatial relations with other objects in its close neighborhood are analyzed.

The assumption is that the flying artificial autonomous agent has to find the concrete building in real urban-type environment. A bitmap image that represents the scene photographed from high altitude is used as the reference representation of the scene. The bitmap is not taken by the agent itself, but it is previously uploaded to the agent's memory and is used by the agent during the scene examination. The agent takes photos of the terrain beneath. Then, representations of single objects as well as the neighbour graph are created. Each single object is compared with the one that is marked as the wanted one on the previously uploaded map. Then, a comparison of the graph fragments is performed that represent the neighbour closest to the analyzed objects pre-selected as identical to the searched one on the created map and the neighbourhood on the object on the uploaded. If the compared neighborhoods are different, then the object is classified as different from the searched one, despite the identical shape. If only one object in the analyzed scene has the same shape and the closest neighborhood as the searched object, it is classified as the searched object. If more than one object in the analyzed scene has the same shape and the closest neighbourhood as the searched object, then in the graph, not only the nearest neighbourhood of a given object is taken into account, but also the nearest neighbours of its nearest neighbours and the graphs obtained in this way are compared. In a such way, the extended neighbour graph is obtained. If necessary, the neighbour graph is extended by the neighbourhoods of subsequent neighbors.

It should be mentioned, that the possibilities of the scene representation go beyond the tools currently used in artificial intelligence. For instance, by rubbing their antennae, ants are able to convey information about the way to the food source, at least in the case when the route has the structure of a branched tree (Ryabko and Reznikova 2009, 2012). It should be stressed that the experiment was designed in such a way that a predefined food route was set up on the water so that the ants could not deviate from it, and after a single scout ant found food, the route was replaced so that the ants could not use the pheromone trail. Other ants were let into the litter box with the scout that returned. After exchanging information, the scout was removed from the litter box, and the ants went straight to the food without fail. This means that the ants can accurately encode the crucial features of the scene although it is not yet known what the nature of this scene representation is. Bees also have specific capabilities related to learning about the environment and transferring information about its structure (Frish 1954; Frish and Lindauer 1956). Namely, bees see in ultraviolet light, and for orientation in space, they use observations of the angle of incidence of sunlight and the polarization of light. Bees use a complex information system an important element of which is the ability to encode information about the location of the food source by means of certain characteristics of the flight of a single bee in the vicinity of the mother hive— the orientation of the circular movements indicates the direction in which the food is located. Although the way of communication of bees is much better studied than the way of communication of ants described above, so far it has not become an inspiration for the development of scene representation algorithms. Similarly, the way of perceiving the environment by dogs in which olfactory stimuli are the basic source of information about the world has not yet been studied in the context of stimuli integration in order to generate an integrated model of the scene.

3. **The analysis of the structure of single objects** is crucial for giving meanings to them (see the next point). In the context of the harvesting robot (Kurpaska et al. 2021), discussed in more detail in the next point, this type of analysis may concern a single fruit in the context of fruits sorting during intelligent harvesting. The strawberry can be mechanically damaged, for example, by biting by a snail. In such a case it changes shape without changing colour. Applying the syntactic contour analysis by using string languages of Jakubowski type (Bielecka 2018; Jakubowski 1990) enabled structural description of damaged and undamaged fruit and, as a consequence, successful qualification (Kurpaska et al. 2021).
4. **Giving meaning to individual objects** involves, among other, assigning specific functionalities to the objects, e.g., whether the analyzed object is a residential building or a defensive bunker, whether a small mound is a natural slight elevation or a large termite mound, etc. Let us analyze the issue more precisely in the context of the vision system of the autonomous robot for harvesting strawberries – see (Kurpaska et al. 2021) for details. In the context of the said robot, the assignment of meanings may, for example, consist in determining whether the detected human figure is a human standing on the robot's collision course, or just a shadow of a human that the robot can overrun. In the context of navigation, it is also important whether the detected obstacle is fixed to the ground or is a light object, e.g., a bucket, that can be moved. The vision system of the robot should perform two various tasks that are mutually independent. First of all, it should ensure navigation on the scene. The robot was supposed to operate in large greenhouses, so it could be assumed that it was a human-made environment. Therefore, the previously developed methodology of scene representation using a neighbourhood graph (Bielecki and Śmigielski 2017) was adapted to represent the scene. The only difference is that the robot operates on limited area. It should be mentioned that the field harvesting robot faces slightly different challenges in the context of the scene representation as well as the trajectory planning and optimizing (Jia et al. 2020). The scene representation is used by the robot to plan its trajectory. An important element is the anti-collision module, which allows for quick recognition of people and animals that the robot may hit while moving. Precise determination of the location of the fruit on the plant and inside is the second task. Furthermore, the analysis of its ripeness phase will have to be done. Furthermore, since the harvesting is assumed to be intelligent, it is necessary to determine whether the fruit is rotten, mouldy, or mechanically damaged. As a consequence, the system should make a decision if the fruit is suitable for collecting or should be discarded due to damage or disease. The scheme of the cognitive vision module of the robot is presented in Fig.1.
5. **The analysis of the relationship between the structural and functional properties of objects and the functional conditions of the agent,** especially in the context of the task being performed and the achievement of the assumed goals, is the next level of the scene understanding. Let us discuss this problem using the example of an autonomous flying agent (Bielecki and Śmigielski 2021). Identifying whether the structure of any examined objects allowed the robot to have collision-free flight under or through the object was one of the tasks performed by the robot. The strategy for the performed test scenario was that the rectangular convex hull is calculated for each side of the representation of the object. Then, the agent verified whether it is possible to navigate through the space

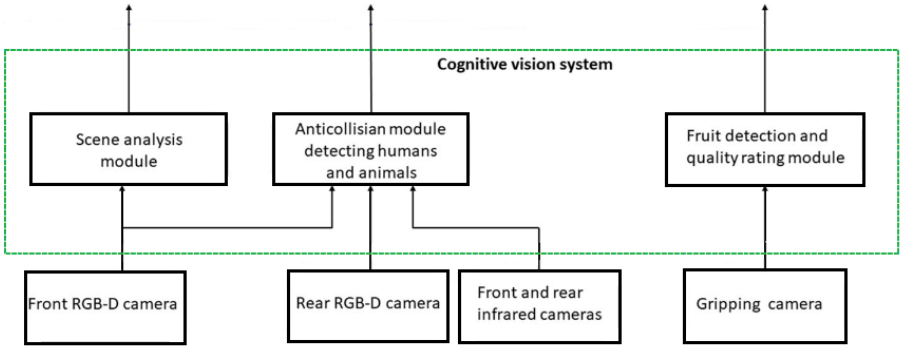

Figure 1. The scheme of the cognitive vision module of the robot for harvesting strawberries (Kurpaska et al. 2021).

bordered by this rectangle without collision with the object. Field trials positively verified the proposed approach.

6. **Analyzing the meaning of the entire scene in the context of the actions performed by the robot is the most general level of cognitive understanding of the scene.** The problem whether the analyzed scene is a fragment of a city after a catastrophe such as an earthquake or whether it is a military training ground for practicing combat tactics in the city is an example of such issue. Without doubt this level of scene understanding, characteristic of humans, is currently unapproachable for autonomous robots.

3.2 Learning process

The functionalities described in subsection 3.1, related to the understanding of the scene by autonomous robots, model the functionalities displayed by humans. Human cognitive abilities, in turn, are the result of the organization of the psyche at the level of cognitive structures. The organization of cognitive structures is known only superficially, and even this aspect that is known, is able to be modelled only in a very simplified way. Nevertheless, models of this type are created and their usefulness in certain applications has been positively verified. One of such models, which was created for the purpose of implementing a mechanism analogous to functional automation, is shown in Fig.2—see Bielecki (2014) for details. The phenomenon of functional automation is analysed as a process as a result of which cognitive structures are replaced by reflexive structures that realize the same task reflexively. Functional automation is a beneficial optimization mechanism when a given activity is performed with high frequency. Thus, when the supervisory module detects that a certain part of the cognitive module is used frequently and the same response pattern is generated as a result, it starts the process of automation. It is implemented in such a way that the reaction generated by the cognitive module, in which knowledge is explicitly encoded, is compared with the reaction generated by the reflex module connected in parallel, which implements the functional stimulus-response relationship. The process is terminated when the output reactions of the reflexive module are equal with sufficient accuracy, to the reactions generated by the cognitive module. Until the automation process is not stopped, the reaction of the

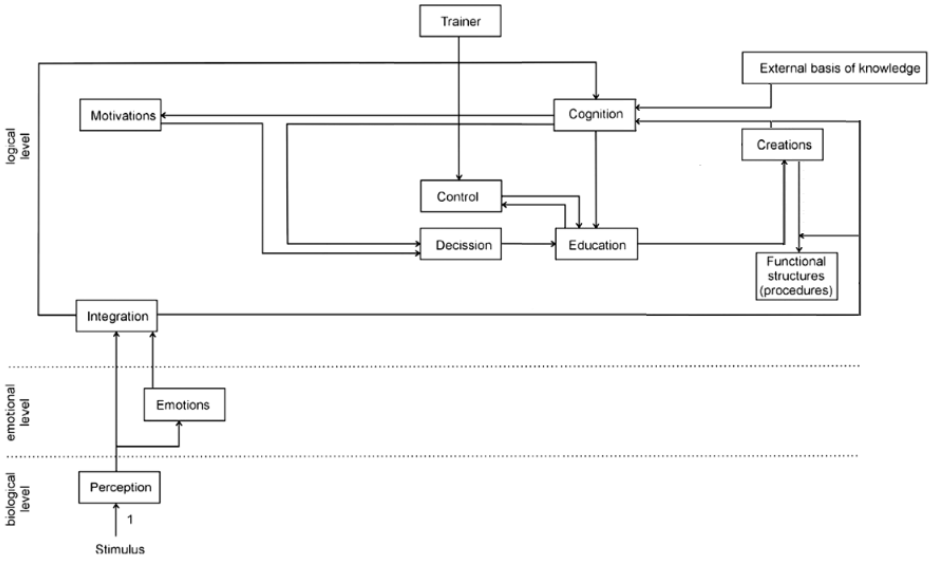

Figure 2. Scheme of the cognitive module of artificial agent. The module is based on the structure of the organization of human cognitive mechanisms related to learning (Bielecki 2014).

cognitive module is the reaction of the whole system— see Fig.3(a). When the learning process is terminated, the cognitive module is disconnected from the reactive loop which means that it does not take part in a direct processing of the input signals—see Fig.3(b). This

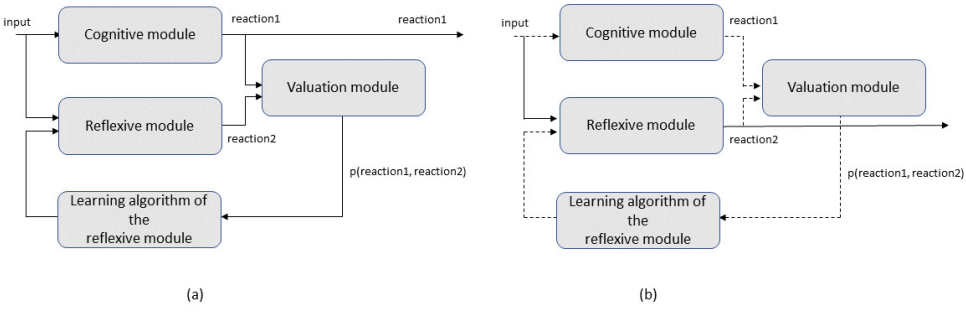

Figure 3. The model of human activity automatization (Bielecki 2014). (a) The agent before automation. (b) The agent after automation.

module becomes the monitoring unit that is activated when reflex responses are no longer satisfactory. Then, the cognitive module is reconnected to the reactive loop and the training of the reflexive module is repeated.

It should be stressed that scene understanding can be considered both in broad and narrow meaning. In the first case, the inner representation of the environment encodes it explicitly, modelling directly, for instance, spatial relations between objects – see (Bielecki

and Śmigielski 2017). Structural and syntactic methods are adequate tools for creating such representations. In the broader sense, however, scene understanding can be considered as all tools that enable the agent to act effectively in the environment. Contemporary AI systems based on computational AI represent the pinnacle of functional automation as defined here. Despite their high operational efficiency, from an epistemological perspective, they remain at the level of signal correlation rather than structural understanding. The absence of explicit syntax in these models precludes the transition from reaction to notion-based cognition, which, in our hierarchy, distinguishes advanced data processing from genuine scene understanding.

4. Epistemic analysis of the robotic cognition abilities

Artificial intelligence, both in general meaning and applied to autonomous robotics, is defined as a feature of artificial systems that enables them to perform tasks that require intelligence (Flasiński 2024). It should be stressed, that intelligence is not understood narrowly, only as a property of the human psyche, but also as natural processes occurring in biological structures and systems that exhibit certain characteristics, such as, for example, the ability to learn, the ability to generalize, and adaptability. In this sense, immune systems, evolutionary processes, insect swarm behaviour, or signal processing in nervous systems, not necessarily related to conscious processes but, for example, to the formation of conditioned reflexes, are also manifestations of intelligence (Flasiński 2016). Cybernetics provides the tools to implement the following scheme for applying this type of natural solutions in artificial intelligence, particularly in autonomous robotics. Namely, the first step is to analyse the biological phenomenon in the context of its functionality. Next, we isolate those aspects that are key to ensuring this functionality. The subsequent step is to create a formal model of the given phenomenon, i.e., the process and biological structures that ensure its implementation. The final step is to implement the formal model in a computer or robotic system. Historically, the biologically inspired AI systems mentioned above were created according to this framework (Bielecki 2024a, 2024b). Let us notice that the mentioned biologically inspired AI systems do not address such concepts as logic, psychology, consciousness, or self-awareness. Nevertheless, they shed light on certain fundamental epistemological issues, for example, those related to aletheology. One of the classic aletheological problems is the problem of truth verification by cognitive entities that have reference to the physical world in which they operate. This is the problem of the correspondence theory of truth, which states that true propositions are those that are consistent with the actual state of affairs in the physical world. However, the correspondence theory of truth, by its very nature, cannot provide a direct algorithm for truth verification. First, we can compare two facts or two propositions, but not a fact with a proposition. Furthermore, significant problems arise in defining correspondence, related, among other things, to the nature of the representation of reality and the manner of expressing this representation. Scene models, created and residing in cognitive modules of autonomous agents, should be verified in terms of their adequacy. This means that for such agents the reference to physical world, considered in the frame of a certain correspondence, plays a crucial role. As a classic philosophical problem, this issue has so far been considered in the context of the human process of learning about the world, including the epistemic procedures of the natural sciences. It seems that the publication (Bielecki 2021) is the first one

in which this problem is considered from the perspective of cybernetics, in particular systems theory, taking into account artificial agents as cognitive subjects. The proposed algorithm of verification of the adequacy of a model is presented in Fig.4. Verification of whether the

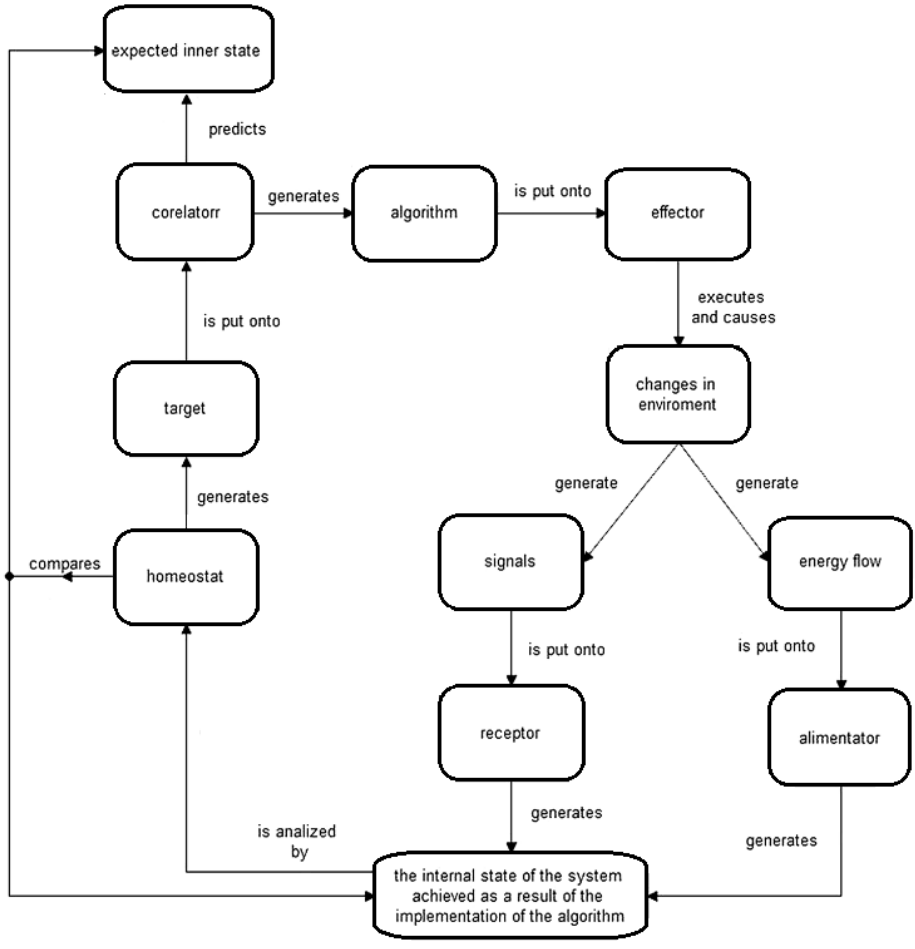

Figure 4. The algorithm of verification of adequacy of a model (Bielecki 2021).

assumed goal has been achieved is the basis of the presented concept. The precise definition of the achievement of the goal is a prerequisite for enabling the aforementioned verification. For an autonomous agent, both biological and artificial, that operates in a physical environment, achieving a goal implies both some change in the environment or the agent's relationship to the environment, as well as the achievement of a certain internal state by the agent. This internal state is predicted by the agent as one of the effects of the successful execution of the algorithm created by the agent. The advantage of the proposed approach is that the internal state achieved by the agent can be directly compared with the state that was predicted as one of the effects of the algorithm execution. On the other hand, the possibility of verifying

changes in the environment is limited by the indirect perception of the environment with the help of sensors in the case of artificial agents and the senses in the case of biological agents. In other words, we do not know to what extent our senses deceive us and we do not know to what extent true information is provided by sensors. The ability to verify the internal state, however, is given to the agent directly. The algorithm presented in Fig.4 can be regarded as a cybernetic rule of inference about the adequacy of the model. Formally, this can be stated as follows—see Bielecki (2021, Section 3). The cognitive subject *S has reason to believe that* ϕ if:

- ϕ is the fragment of the model such that there exists at least one goal, there exist conditions, for instance a range of parameters, and there exists the algorithm created on the basis of the ϕ such that the goal is achievable by using the algorithm provided that the conditions are satisfied;
- ϕ is encoded in the cognitive module of the autonomous agent *S*;
- *S* performed the algorithm and, as a result, *S* achieved its goal, which manifests, among others, by achieving the inner state the agent assumed before performing of the algorithm.

It should be emphasized that while the *cybernetic truth verification loop* is presented in a formalized manner, its individual components are grounded in established paradigms of visual information processing in autonomous systems. The implementation of the scene understanding level utilizes the structure of neighbourhood graphs, which has been previously validated as an effective tool for the representation and recognition of objects within complex urban environments (Bielecki and Śmigielski 2017). Furthermore, the operational efficiency of the system in unstructured settings is supported by outdoor experimental data obtained from unmanned flying agents using monocular vision, where algorithms successfully performed three-dimensional analysis of structures under varying lighting and geometric conditions (Bielecki and Śmigielski 2021). In the present work these methods are theoretically analysed by introducing a rigorous truth verification mechanism that integrates raw sensory data with higher levels of cognitive reasoning. In a such way, the above cybernetic approach sheds new light on the problem of truth verification.

Turning to the problem of the limitations of cognitive capabilities of the autonomous robots, it should be emphasized that it is difficult to find a valuable philosophical analysis of this issue. There appear to be no methodological limitations to adapting biological mechanisms for AI and autonomous robotics. There are no fundamental problems in researching them, creating their models and implementing them. It should be also noted that both AI systems themselves and their learning algorithms can be biologically inspired—see, for example, the functional automation mechanism described in the previous chapter. It should be mentioned that these topics are included in biomimetics, called also biomimicry or bionics, that is an intensively studied branch of science (Bhushan 2009).

When it comes to psychologically inspired AI systems, however, the situation is different. A thorough discussion of the problem in the light of Aristotle, St. Thomas Aquinas and, first of all, Kant's approach was conducted in a series of works by Flasiński (1997, 2016, 2024). Let us sum up briefly his analysis. Namely, according to the philosophy of mind formulated by Kant (1838), the intellect is carefully distinguished from the reason. According to Kant, the intellect realizes the operations of the logical type, establishing relationships between phenomena and creating categories. Reason, on the other hand, is capable of

creating transcendental ideas and primary principles on the basis of which greater premises are created in syllogistic reasoning.

Some, but not all, functions of the intellect can currently be implemented by artificial intelligence systems. In addition to the already mentioned logical inference, artificial intelligence systems are able to group objects into categories based on their characteristics, but they are not able to create concepts corresponding to these categories. For example, they can group round objects into one collection, but they will not create the idea of the circle. In general, reasoning is satisfactorily realized by AI systems due to the development of the contemporary mathematical logic that is fully formalized and, as a consequence, algorithmized.

Pronouncing judgements, however, can be partially modelled by using knowledge stored in knowledge bases or by realization of inductive learning based on examples. Such schema is realized, for instance, in artificial neural networks (Bielecki 2024a). Nevertheless, since the contemporary psychology does not describe the mechanism of conceptualisation, this cognitive functionality is unattainable for contemporary AI systems. As a consequence, formulation of notions, propositions and judgements remains beyond capabilities of AI systems. Arguably, consciousness and self-awareness are also unattainable for AI systems.

It should be noted that the claim regarding the inability of artificial systems to conceptualize, as raised in this paper, may be perceived differently from a functionalist perspective. From the standpoint of functionalism, cognitive processes can be realized through mechanisms entirely distinct from human psychology, provided they lead to equivalent results. However, within the context of cognitive epistemology based on Kantian theory, a crucial distinction must be made between functional categorization and essential conceptualization (Flasiński 1997, 2024). While AI effectively groups objects into sets based on features, which externally may resemble the formation of concepts, the absence of a formalized psychological theory concerning the emergence of ideas prevents the implementation of a mechanism that would grant these categories the status of abstract concepts and judgements. Thus, this barrier is not merely technical but stems from fundamental differences in how formal systems and living cognitive subjects constitute meaning.

5. Concluding remarks and comments

In this paper, we discussed the problem of representation and analysis of the scene and building its models in the context of cognitive possibilities of future, fully autonomous, artificial agents, primarily robots. With all the current, considerable possibilities in the field of implementing cognitive modules in artificial intelligence systems and with the perspectives of development of such systems, the natural question arises: "Do such systems have significant barriers compared to human capabilities?" The analysis of this problem goes beyond the framework of computer science and cybernetics, at least at the current stage of development of these sciences, entering the area of philosophy. At the current stage of artificial intelligence development in general and the problem of scene understanding in particular, it seems that there are not fundamental barriers to development of biologically inspired AI systems. Furthermore, they can shed light on some fundamental epistemological problems, such as truth verification. However, the problem of the possibility of constructing thinking or conscious machines remains open (Bremer and Flasiński 2022). Perhaps the consciousness and self-awareness are these barriers, which by their very nature artificial systems cannot cross. The distinction between advanced

biological mimicry and true consciousness is often perceived as "porous". A rigorous analysis of formal foundations of AI systems suggests, however, a fundamental categorical barrier. This limitation is rooted in three primary areas. First, as noted by Flasiński (Flasiński 1997, 2024), current AI paradigms lack formalized models for the basic cognitive operation of the autonomous creation of concepts (*indivisibilium intelligentia*). While AI excels at reasoning (*ratiocinare*) through logic and pattern recognition through induction, it cannot perform the qualitative leap of abstraction required for true conceptualization, as this process has not been operationalized in a way that allows for algorithmic implementation.

Second, the perceived "porosity" of the boundary between machine performance and human reason is frequently a result of a misuse of language (Bremer and Flasiński 2022). Following the analytic tradition of Wittgenstein and Chomsky, it is argued that ascribing mental qualities like "self-awareness" to a system based on its behavioral output (e.g., passing a Turing-like test) constitutes a category mistake. AI demonstrates performance without competence; it simulates the results of consciousness without possessing the underlying mental states.

Finally, the transition from weak AI to strong AI remains blocked by the gap between syntax and semantics. As it is emphasized (Flasiński 2024), the computational nature of AI, whether symbolic or connectionist, is limited to the manipulation of representations according to formal rules. Without a biological or ontological foundation for understanding, the system remains an imitation of intelligence. Thus, the barrier to AI consciousness is not a temporary technological hurdle but a structural limitation of the current computational paradigm.

To sum up, the analysis conducted in this paper demonstrates that the problem of scene understanding is not merely a technological challenge, but a profound theoretical inquiry into the nature of cognition and epistemology in artificial systems. By situating robotic vision within the framework of cybernetics and Kantian philosophy, we have shown that while biological inspirations provide a robust path for developing functional autonomy, they do not automatically resolve the fundamental limitations related to conceptualization. The technological solutions described herein serve as a practical verification of theoretical assertions regarding the correspondence theory of truth and the adequacy of world models. Ultimately, the distinction between intellect, i.e., capable of logical operations and categorization, and reason, i.e. capable of creating transcendental ideas, remains the primary theoretical boundary that separates current artificial intelligence from human-like cognitive subjects.

It should also be mentioned that a very important and often overlooked problem is the impact of emotions on the learning process and the problem of emotions as a learning mechanism. These issues are also considered in the context of artificial intelligence and cognitive abilities of artificial systems (Abbate 2023; Assunção et al. 2022; Bielecki, Bielecka, and Bielecki 2017; Rutar, Wiese, and Kwisthout 2022). This topic, however, goes beyond the scope of this article.

References

Abbate, Francesco. 2023. Natural and Artificial Intelligence: A Comparative Analysis of Cognitive Aspects. *Minds and Machines* 33 (4): 791–815. https://doi.org/10.1007/s11023-023-09646-w.

Assunção, Gustavo, Bruno Patrão, Miguel Castelo-Branco, and Paulo Menezes. 2022. An Overview of Emotion in Artificial Intelligence. *IEEE Transactions on Artificial Intelligence* 3 (6): 867–886. https://doi.org/10.1109/TAI.2022.3159614.

Bajcsy, Ruzena, Yiannis Aloimonos, and John K. Tsotsos. 2018. Revisiting active perception. *Autonomous Robots* 42 (2): 177–196. https://doi.org/10.1007/s10514-017-9615-3.

Barreiro, C., Jose M. Barreiro, J. A. Lara, D. Lizcano, M. A. Martínez, and J. Pazos. 2020. The Third Construct of the Universe: Information. *Foundations of Science* 25 (2): 425–440. https://doi.org/10.1007/s10699-019-09630-7.

Bhushan, Bharat. 2009. Biomimetics: lessons from nature–an overview. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 367 (1893): 1445–1486. https://doi.org/10.1098/rsta.2009.0011.

Bielecka, Marzena. 2018. Syntactic-geometric-fuzzy hierarchical classifier of contours with application to analysis of bone contours in X-ray images. *Applied Soft Computing* 69:368–380. https://doi.org/10.1016/j.asoc.2018.04.038.

Bielecka, Marzena, Andrzej Bielecki, Aleksander Suchorab, and Igor Wojnicki. 2025. Hierarchical Structural Information – Theory and Applications. In *Computational Science – ICCS 2025. 25th international conference: Singapore, July 7–9, 2025: proceedings, Pt. 3,* edited by Michael H. Lees, Wentong Cai, Siew Ann Cheong, Yi Su, David Abramson, Jack J. Dongarra, and Peter M. A. Sloot, 15905:48–59. Series Title: Lecture Notes in Computer Science. Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-97632-2_4.

Bielecki, Andrzej. 2014. A Model of Human Activity Automatization as a Basis of Artificial Intelligence Systems. *IEEE Transactions on Autonomous Mental Development* 6 (3): 169–182. https://doi.org/10.1109/TAMD.2014.2319740.

Bielecki, Andrzej. 2021. The Systemic Concept of Contextual Truth. *Foundations of Science* 26 (4): 807–824. https://doi.org/10.1007/s10699-020-09713-w.

Bielecki, Andrzej. 2024a. Artificial Neural Networks. In *Cognitive Science. Vol. 13,* edited by Józef Bremer, 299–314. Social Dictionaries. Kraków: Ignatianum University Press.

Bielecki, Andrzej. 2024b. Cybernetics. In *Cognitive Science. Vol. 13,* edited by Józef Bremer, 281–298. Social Dictionaries. Kraków: Ignatianum University Press.

Bielecki, Andrzej, Marzena Bielecka, and Przemysław Bielecki. 2017. Conditioned Anxiety Mechanism as a Basis for a Procedure of Control Module of an Autonomous Robot. In *Artificial Intelligence and Soft Computing. 16th International Conference: ICAISC 2017 Zakopane, Poland, June 11–15, 2017. Proceedings, Pt. 2,* edited by Leszek Rutkowski, Marcin Korytkowski, Rafał Scherer, Ryszard Tadeusiewicz, Lotfi A. Zadeh, and Jacek M. Zurada, 10246:390–398. Series Title: Lecture Notes in Computer Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-59060-8_35.

Bielecki, Andrzej, Tomasz Buratowski, M. Ciszewski, and Piotr Śmigielski. 2016. Vision based techniques of 3D obstacle reconfiguration for the outdoor drilling mobile robot. In *Artificial intelligence and soft computing. lecture notes in computer science,* edited by Leszek Rutkowski, Marcin Korytkowski, Rafał Scherer, Ryszard Tadeusiewicz, Lotfi A. Zadeh, and Jacek M. Zurada, 9693:602–612. Series Title: Lecture Notes in Computer Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-39384-1_53.

Bielecki, Andrzej, Tomasz Buratowski, and Piotr Śmigielski. 2013. Recognition of two-dimensional representation of urban environment for autonomous flying agents. *Expert Systems with Applications* 40 (9): 3623–3633. https://doi.org/10.1016/j.eswa.2012.12.068.

Bielecki, Andrzej, Tomasz Buratowski, and Piotr Śmigielski. 2014. Three-Dimensional Urban-Type Scene Representation in Vision System of Unmanned Flying Vehicles. In *Artificial Intelligence and Soft Computing. 13th International Conference, ICAISC 2014, Zakopane, Poland, June 1-5, 2014, Proceedings, Part I,* edited by David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Alfred Kobsa, Friedemann Mattern, John C. Mitchell, et al., 8467:662–671. Series Title: Lecture Notes in Computer Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-07173-2_56.

Bielecki, Andrzej, and Michael Schmittel. 2022. The Information Encoded in Structures: Theory and Application to Molecular Cybernetics. *Foundations of Science* 27 (4): 1327–1345. https://doi.org/10.1007/s10699-022-09830-8.

Bielecki, Andrzej, and Piotr Śmigielski. 2017. Graph representation for two-dimensional scene understanding by the cognitive vision module. *International Journal of Advanced Robotic Systems* 14 (1): 1729881416682694. https://doi.org/10.1177/1729881416682694.

Bielecki, Andrzej, and Piotr Śmigielski. 2021. Three-Dimensional Outdoor Analysis of Single Synthetic Building Structures by an Unmanned Flying Agent Using Monocular Vision. *Sensors* 21:s21217270. https://doi.org/10.3390/s21217270.

Bielecki, Andrzej, and Ryszard Stocki. 2023. The concept of structural information and possible applications. *Philosophical Problems in Science (Zagadnienia Filozoficzne w Nauce),* no. 75, 157–183. https://doi.org/10.59203/zfn.75.654.

Bremer, Józef, and Mariusz Flasiński. 2022. The Turing Test, or a Misuse of Language when Ascribing Mental Qualities to Machines. *Forum Philosophicum* 27 (1): 6–25. https://doi.org/10.35765/forphil.2022.2701.01.

Chaib-draa, B. 2002. Causal maps: theory, implementation, and practical applications in multiagent environments. *IEEE Transactions on Knowledge and Data Engineering* 14 (6): 1201–1217. https://doi.org/10.1109/TKDE.2002.1047761.

Flasiński, Mariusz. 1997. "Every Man in His Notions" or Alchemists' Discussion on Artificial Intelligence. *Foundations of Science* 2 (1): 107–121. https://doi.org/10.1023/A:1009687513096.

Flasiński, Mariusz. 2016. *Introduction to Artificial Intelligence.* Cham: Springer Nature.

Flasiński, Mariusz. 2019. *Syntactic Pattern Recognition.* New Jersey-London-Singapore: World Scientific.

Flasiński, Mariusz. 2024. Artificial Inteligence. In *Cognitive Science. Vol. 13,* edited by Józef Bremer, 315–330. Social Dictionaries. Kraków: Ignatianum University Press.

Floreano, Dario, and Robert J. Wood. 2015. Science, technology and the future of small autonomous drones. *Nature* 521 (7553): 460–466. https://doi.org/10.1038/nature14542.

Frish, Karl. 1954. *The Dancing Bees.* Berlin-Gotingen-Heidelberg: Springer.

Frish, Karl, and Martin Lindauer. 1956. The "Language" and Orientation of the Honey Bee. *Annual Review of Entomology* 1 (1): 45–58. https://doi.org/10.1146/annurev.en.01.010156.000401.

Friston, Karl. 2009. The free-energy principle: a rough guide to the brain? *Trends in Cognitive Sciences* 13 (7): 293–301. https://doi.org/10.1016/j.tics.2009.04.005.

Hu, Cheng, Farshad Arvin, Caihua Xiong, and Shigang Yue. 2017. Bio-inspired embedded vision system for autonomous micro-robots: The LGMD case. *IEEE Transactions on Cognitive and Developmental Systems* 9:241–254.

Jakubowski, Ryszard. 1990. Decomposition of complex shapes for their structural recognition. *Information Sciences* 50 (1): 35–71. https://doi.org/10.1016/0020-0255(90)90004-T.

Jia, Weikuan, Yan Zhang, Jian Lian, Yuanjie Zheng, Dean Zhao, and Chengjiang Li. 2020. Apple harvesting robot under information technology: A review. *International Journal of Advanced Robotic Systems* 17 (3): 1729881420925310. https://doi.org/10.1177/1729881420925310.

Kang, Soo-Han, and Ji-Hyeong Han. 2023. Video Captioning Based on Both Egocentric and Exocentric Views of Robot Vision for Human-Robot Interaction. *International Journal of Social Robotics* 15 (4): 631–641. https://doi.org/10.1007/s12369-021-00842-1.

Kant, Immanuel. 1838. *Critick of Pure Reason.* Translated by Francis Haywood. London: William Pickering.

Katsev, Max, Anna Yershova, Benjamín Tovar, Robert Ghrist, and Steven M LaValle. 2011. Mapping and Pursuit-Evasion Strategies For a Simple Wall-Following Robot. *IEEE Transactions on Robotics* 27 (1): 113–128. https://doi.org/10.1109/TRO.2010.2095570.

Krzanowski, Roman. 2020. What Is Physical Information? *Philosophies* 5 (2): 10. https://doi.org/10.3390/philosophies5020010.

Krzanowski, Roman. 2022. *Ontological information: information in the physical world.* World Scientific series in information studies, vol. 13. Singapore: World Scientific.

Krzanowski, Roman, and Pawel Polak. 2022. The meta-ontology of AI systems with human-level intelligence. *Philosophical Problems in Science (Zagadnienia Filozoficzne w Nauce),* no. 73, 197–230. https://doi.org/10.59203/zfn.73.610.

Kurpaska, Sławomir, Andrzej Bielecki, Zygmunt Sobol, Marzena Bielecka, Magdalena Habrat, and Piotr Śmigielski. 2021. The Concept of the Constructional Solution of the Working Section of a Robot for Harvesting Strawberries. *Sensors* 21 (11): 3933. https://doi.org/10.3390/s21113933.

Li, Jun, Xue Mei, Danil Prokhorov, and Dacheng Tao. 2017. Deep Neural Network for Structural Prediction and Lane Detection in Traffic Scene. *IEEE Transactions on Neural Networks and Learning Systems* 28 (3): 690–703. https://doi.org/10.1109/TNNLS.2016.2522428.

Polak, Paweł. 2023. Philosophy in technology: A research program. *Philosophical Problems in Science (Zagadnienia Filozoficzne w Nauce),* no. 75, 59–81. https://doi.org/10.59203/zfn.75.682.

Quintas, João. 2023. Improving Visual Perception of Artificial Social Companions Using a Standardized Knowledge Representation in a Human-Machine Interaction Framework. *International Journal of Social Robotics* 15 (3): 425–444. https://doi.org/10.1007/s12369-021-00859-6.

Rosslenbroich, Bernd. 2014. *On the Origin of Autonomy: A New Look at the Major Transitions in Evolution.* Cham: Springer International Publishing AG.

Roterman, Irena, and Leszek Konieczny. 2023. Protein Is an Intelligent Micelle. *Entropy* 25 (6): 850. https://doi.org/10.3390/e25060850.

Rutar, Danaja, Wanja Wiese, and Johan Kwisthout. 2022. From representations in predictive processing to degrees of representational features. *Minds and Machines* 32 (3): 461–484. https://doi.org/10.1007/s11023-022-09599-6.

Ryabko, Boris, and Zhanna Reznikova. 2009. The Use of Ideas of Information Theory for Studying "Language" and Intelligence in Ants. *Entropy* 11 (4): 836–853. https://doi.org/10.3390/e11040836.

Ryabko, Boris, and Zhanna Reznikova. 2012. Ants and bits. *IEEE Information Theory Society Newsletter* 62 (5): 17–20.

Scott, Alwyn. 1995. *Stairway to the Mind: The Controversial New Science of Consciousness.* New York: Springer New York.

Siagian, Christian, and Laurent Itti. 2009. Biologically Inspired Mobile Robot Vision Localization. *IEEE Transactions on Robotics* 25 (4): 861–873. https://doi.org/10.1109/TRO.2009.2022424.

Tadeusiewicz, Ryszard. 1992. *Vision Systems of Industrial Robots.* Warszawa: WNT Press.

Tadeusiewicz, Ryszard. 2010. New trends in neurocybernetics. *Computer Methods in Materials Science* 10 (1): 1–7. https://doi.org/10.7494/cmms.2010.1.0271.

Tejera, Gonzalo, Martin Llofriu, Alejandra Barrera, and Alfredo Weitzenfeld. 2018. Bio-Inspired Robotics: A Spatial Cognition Model integrating Place Cells, Grid Cells and Head Direction Cells. *Journal of Intelligent & Robotic Systems* 91 (1): 85–99. https://doi.org/10.1007/s10846-018-0852-2.

Umbrico, Alessandro, Riccardo De Benedictis, Francesca Fracasso, Amedeo Cesta, Andrea Orlandini, and Gabriella Cortellessa. 2023. A Mind-inspired Architecture for Adaptive HRI. *International Journal of Social Robotics* 15 (3): 371–391. https://doi.org/10.1007/s12369-022-00897-8.

Wang, Heng, Bin Wang, Bingbing Liu, Xiaoli Meng, and Guanghong Yang. 2017. Pedestrian recognition and tracking using 3D LiDAR for autonomous vehicle. *Robotics and Autonomous Systems* 88:71–78. https://doi.org/10.1016/j.robot.2016.11.014.

Philosophical Problems in Science (2025), **79**, 21–39
(Zagadnienia Filozoficzne w Nauce)
doi:10.59203/zfn.79.729

ARTICLE

From Digital Twins to Entangled Twins: A phenomenological model for Industry 5.0

Floriana Ferro,[*†] Alessandro Giuseppe Privitera,[‡] Andrea Gulli,[¶] and Michele Geronazzo[¶]

†University of Trieste, Italy
‡University of Udine, Italy
¶University of Padova, Italy
[*]Corresponding author. Email: florianagmferro@gmail.com

Abstract

This article introduces the concept of Entangled Twins (ETs) as an interdisciplinary framework emerging from a sustained dialogue between Human–Computer Interaction (HCI) research and phenomenological philosophy. Building on the Digital Twin (DT) paradigm while critically reworking its representational logic, ETs are conceived as relational configurations in which human users and AI-driven systems become dynamically entangled, with significant implications for agency, responsibility, and human-centered design. The theoretical framework draws on phenomenological analyses of intersubjectivity and embodiment (Husserl, Merleau-Ponty, and Sartre) together with design-oriented concerns in HCI. Within this perspective, human–machine interaction is conceptualized in terms of a We-subject: a plural and co-constituted configuration of agency. Through this interdisciplinary approach, a prototypical audio-augmented experience is examined as a phenomenologically informed case, elucidating the conditions under which human–machine entanglement can emerge. On this basis, the paper argues that ETs raise distinctive questions concerning responsibility and agency, articulating an ethics of entanglement capable of informing the human-centered aspirations of Industry 5.0.

Keywords: entangled twins, digital twin, human-computer interaction, phenomenology, ethics

Introduction

In recent years, digital technologies have profoundly reshaped the ways in which humans interact with complex systems, environments, and decision-making processes. Among these technologies, the concept of the *Digital Twin* (DT) has gained significant prominence across industrial, scientific, and organizational domains. Originally developed as a tool for simulation and optimization in aerospace engineering contexts (Glaessgen and Stargel 2012), the DT has progressively expanded its scope, entering domains where human presence, perception, and action are no longer peripheral but central to the system's functioning: DTs have become widely used in contexts where physical and digital processes are mirrored and coordinated in real time (Barricelli, Casiraghi, and Fogli 2019). Their integration into Mixed Reality

(MR) environments—hybrid spaces where physical and digital components interact—has further broadened their scope, raising fundamental questions about identity, agency, and responsibility in technologically mediated experience (Jerald 2016).

As DTs are deployed in increasingly immersive and interactive environments, their traditional understanding as representational or predictive models reveals important limitations. In contexts where users do not merely consult a digital system but engage with it in real time—through embodied action, perceptual feedback, and adaptive interaction—the relationship between humans and digital systems can no longer be adequately described in terms of mirroring, modeling, or data-driven optimization alone. What emerges instead is a relational configuration in which human experience and technological behavior are dynamically co-shaped.

This paper advances the concept of *Entangled Twins* (ETs) (Privitera and Geronazzo 2024) articulating it as a theoretical and operational development of the DT paradigm. Rather than conceiving the digital system as a detached representation of a physical or human process, ETs shift the emphasis from simulation to entanglement: a dynamic, reciprocal relationship between human users and AI-driven technical systems. The shift from DT to ETs does not involve attributing human subjectivity to machines, nor does it reduce human agency to computational processes. Instead, it reflects a posthuman perspective, according to which agency and meaning emerge within situated relations between humans, technologies, and environments.

The contribution of this paper is threefold. First, it offers a clear conceptual distinction between DTs and ETs, supported by a phenomenologically informed framework (Husserl 1960; Merleau-Ponty 1968; Sartre 1992) that clarifies the nature of the human–machine relation at stake. Second, it introduces an operational translation of this framework through Human-Computer Interaction (HCI) research (Geronazzo 2022; Privitera and Geronazzo 2024), designed to support the analysis and design of interactive systems in which relational dynamics play a central role. It also presents a prototypical case study in a MR setting, not as a statistical validation, but as an exploratory investigation of how entangled human–machine relations can emerge, stabilize, or fail in practice. Methodologically, the paper adopts an interdisciplinary approach that integrates phenomenological analysis with HCI research. The study does not aim to offer a comprehensive empirical generalization, but rather to articulate the structural conditions under which ETs can be meaningfully described and ethically assessed. The third contribution consists in addressing ethical implications of this model, so that ETs are discussed in relation to current debates surrounding Industry 5.0 (Alves, Lima, and Gaspar 2023).

The paper is structured as follows. The first section introduces and compares DTs and ETs, grounding the distinction in a phenomenological account of relational experience. The second section presents an operational framework that translates this philosophical perspective into design-relevant terms, introduces the notion of prototypical experience, and discusses a case study as an example of an entangled human–machine interaction in a MR environment. The third section reflects on the ethical implications of the ETs model, focusing on relational agency and responsibility in contemporary technological systems.

1. From Digital Twins to Entangled Twins

1.1 Digital Twins and Entangled Twins: Definitions and Conceptual Distinctions

Recent developments in HCI and Artificial Intelligence (AI) have given rise to a proliferation of concepts related to so-called "digital twins," including Virtual Twins, Predictive Twins, and Hybrid Twins, each aiming to capture the evolving forms of interaction between humans and their digital counterparts (Abburu et al. 2020). Despite this conceptual richness, no shared theoretical framework has yet emerged that adequately clarifies the experiential and relational implications of these models. In this study, we focus on the notion of the DT, refining it in view of human–machine entanglement (Frauenberger 2019).

For the purposes of this paper, a DT can be defined as a digital model of a physical or socio-technical system, a sort of digital mirror that is continuously updated through data, allowing the system's state and behavior to be monitored, simulated, and anticipated over time. In contemporary applications, DTs increasingly incorporate machine-learning techniques, allowing them to adapt to new inputs and support complex decision-making processes (Madni, Madni, and Lucero 2019).

Yet, even when endowed with adaptive and semi-autonomous capabilities, the DT paradigm remains fundamentally representational. The DT primarily serves as a model to be consulted, interpreted, and acted upon by a human decision-maker who remains structurally external to the system. This configuration becomes conceptually insufficient in immersive and interactive contexts, where users do not simply observe or interrogate a model but engage with digital systems through embodied action, perception, and continuous feedback. In such cases, the DT framework struggles to account for the experiential, affective, and relational dimensions of interaction.

The limits of the DT paradigm become particularly evident when examining the concepts of "artificial" and "intelligence" that underpin it. Rather than treating the artificial as the opposite of the natural, we follow (Hayles 1999) in understanding embodiment and cognition as processes that can span biological and technological substrates. From this perspective, artificial systems are not external tools but extensions and transformations of human intentionality, embedded within broader material and relational contexts. Similarly, intelligence cannot be reduced to symbolic computation or algorithmic optimization. As (Floridi 2010) argues, in the age of the infosphere and digital re-ontologization, intelligent behavior emerges within networks of information, interaction, and meaning. Intelligence, in this sense, is enacted rather than possessed: it takes shape through the coupling of algorithmic systems with physical processes, sensor data, and human engagement. This shift is crucial for understanding ETs, since their agency does not reside in the digital system alone, but emerges within ongoing human–machine coupling.

This relational and situated understanding implies that a DT is not merely a passive representational model but an active component within decision-making ecologies. However, its agency remains indirect and instrumental. This reconceptualization paves the way for the emergence of ETs: AI-augmented systems that no longer merely mirror physical or human processes but participate in the constitution of experience, orientation, and shared meaning.

ETs are not conceived as digital *Doppelgängers*—mere replicas of human empirical selves—or artificial autonomous subjects. Our interpretation aligns with the notion of the extended and embodied mind, as developed by (Clark and Chalmers 1998) and further elaborated by (Gallagher and Zahavi 2020), in which the human mind is not localized in the brain

or the body alone, but is dynamically coupled with external structures and environmental affordances. According to Clark and Chalmers the mind is fundamentally shaped by what they call "active externalism," wherein cognitive processes are extended beyond the skin and skull, into tools, symbolic media, and digital systems (Clark and Chalmers 1998, p.7). The boundaries of cognition are thus porous, contingent on the coupling with devices and environments that scaffold and co-constitute our mental life. From this point of view, ETs are entangled cognitive configurations in which human and machine collaborate in shaping experience, meaning, and behavior. This shift moves us away from the notion of passive simulation and toward a framework of co-evolution and shared intentionality.

The transition from DTs to ETs does not involve attributing human subjectivity or consciousness to machines, nor does it reduce human agency to computational processes. Rather, ETs embody a posthuman but non-reductionist perspective, which avoids reducing human experience to algorithmic computation while also resisting the attribution of human subjectivity to machines. Within this perspective, agency and meaning emerge from situated relations between humans, technologies, and environments.

The conceptual shift proposed here can be summarized in the following table:

Table 1. Comparative table of DTs and ETs

Dimension	Digital Twins (DTs)	Entangled Twins (ETs)
Core function	Simulation and prediction	Co-adaptive interaction
Ontological orientation	Representational	Relational
Human-system relation	External and mediated	Situated and performative
Role of AI	Optimization and modeling	Co-regulation and adaptation
Agency	Indirect and instrumental	Relative and distributed
Experiential focus	Decision support	Meaning-making and orientation

This comparison highlights that ETs do not represent an incremental technical improvement over DTs, but a reorientation of the human–machine relationship itself.

1.2 Phenomenological Foundations: Intersubjectivity, Embodiment and the We-subject

To articulate the theoretical foundations of ETs, a phenomenological account of experience and relation is required. Phenomenology understands experience not as an internal mental state but as a structured relation between a subject and its world. From this standpoint, relationality is not secondary but constitutive of experience.

In Husserlian phenomenology, intersubjectivity is a transcendental condition for the constitution of experience and meaning. Self-awareness itself is already co-constituted through the presence of the Other as another subjectivity—an *alter ego* whose experiential field is both similar to and irreducibly distinct from mine (Husserl 1960). This insight is developed further by Alfred Schutz (1967), Edith Stein (1989), Max Scheler (2008), and Jean-Paul Sartre (1992). All these authors emphasize that intersubjective relations precede reflective thematization. That is, I am always already situated in a network of social and affective relations before I become explicitly aware of them. As Zahavi (2015) argues, the presence of others is given pre-reflectively, as an immediate structure of experience rather than as an inferential achievement. In the context of ETs, this implies that digital agents must be more than interactive interfaces: they must be capable of participating in pre-reflective

relationality, aligning themselves with the user's affective rhythms, intentions, and embodied habits.

A particularly relevant conceptual resource for articulating this relational configuration is Sartre's notion of the "We-subject". In *Being and Nothingness*, Sartre distinguishes between the Us-object and the We-subject. The former designates a collectivity defined from an external point of view, a passive aggregation of individuals objectified by a third party. The latter, by contrast, emerges through the convergence of intentionalities oriented toward a common situation or project. It is neither a fusion nor a simple aggregation of individual egos, but a relational configuration that preserves alterity while enabling shared orientation (Sartre 1992, p.413). This distinction is of critical importance in the case of ETs. Digital systems risk being reduced to mere tools, unless their design and functioning allow for a more complex, participatory role in shaping experience. If we take seriously the phenomenological claim that subjectivity is co-constituted through relational structures, then ETs must be configured not as objects to be used, but as co-agents in the constitution of a We-subject configuration. This should not be understood as attributing a fully constituted subjectivity to the digital component of ETs. Instead, it points to an extension of human subjectivity through forms of digital agency that remain structurally entangled with human agency. The implications of this shift are both ontological and ethical, as they require a reconsideration of the locus of responsibility and the distribution of agency within human–machine relations.

These relational dynamics can be further clarified by considering the role of embodiment. From a phenomenological perspective, perception and action arise through a reciprocal interplay between bodily intentionality and environmental affordances. Within contemporary digital environments, the world as lived (*Umwelt*) can no longer be understood as purely natural. Pervasive digitalization and the rise of AI-based systems have destabilized traditional distinctions between organic and artificial, real and virtual (Floridi 2010). This ontological shift provides the background for understanding ETs not merely as tools, but as participants in shared spaces of agency, perception, and meaning-making.

Relational and posthuman ontologies such as Karen Barad's agential realism (Barad 2007) and Bruno Latour's Actor–Network Theory (Latour 2005) have contributed to highlighting the distributed nature of agency across human and non-human actors. These approaches resonate with the perspective developed here, though they do not replace the phenomenological account grounded in embodiment and perception. In this respect, Merleau-Ponty's notion of flesh (*chair*) plays a central role. In *The Visible and the Invisible* (Merleau-Ponty 1968), he introduces flesh as the ontological element that underlies both perception and being, dissolving the rigid distinction between subject and object. Flesh is not matter, nor is it mind; it is the reversible medium through which the perceiving and the perceived, the human and the world, are intertwined. Applying this concept to the ETs framework, we propose that the human and the machine are not external to one another but emerge from a shared ontological field—a field of perceptual, affective, and technological entanglement.

Within this phenomenological horizon, the notion of *Sinngebung* (sense-bestowal) can be reconsidered. In Husserl's account, sense-bestowal is not a purely subjective act, but a process through which meaning is constituted within experience (Husserl 1983). In the context of ETs, *Sinngebung* cannot be understood as an exclusively individual operation, nor as a function delegated to the technical system. Rather, it emerges as a relational process involving users, digital systems, environments, and design choices. This distributed sense-making is evident

in a range of mixed-reality applications, from learning environments to industrial simulations (Degli Innocenti et al. 2019; Geronazzo and Serafin 2023a; Wang, Shen, and Lee 2025), where experience is not merely mediated by technology but constituted through it.

Taken together, these considerations provide the phenomenological foundation for ETs as relational configurations of human–machine interaction. ETs are not artificial subjects, nor mere tools, but situated participants in ontologically and ethically charged fields of experience. This framework grounds the operational model introduced in the next section, where these concepts are translated into design-relevant terms.

2. Translating Phenomenology into a Prototypical Case

2.1 The ICE Framework as an Operational Translation

The phenomenological framework developed in the previous section calls for an operational articulation capable of addressing human–machine interaction without reducing it to purely technical performance or behavioral outcomes. To this end, we introduce the ICE framework (Geronazzo 2022), which functions as a phenomenologically informed interpretive model for tracing relational dynamics in MR experiences.

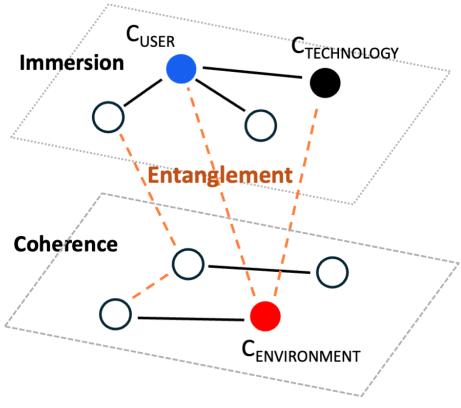

Figure 1. The multilayer interconnected network of actors in human–machine entangled experiences. Colored spheres identify example configurations (C) of user, technology, and environment.

Figure 1 schematically represents the relational configuration underlying ICE, where user, technology, and environment form a dynamic network. ICE offers a set of experiential dimensions through which the conditions of possibility for relational entanglement can be described. For this reason, we speak of *traces* or *footprints* rather than measurements, emphasizing that entanglement cannot be exhaustively captured through numerical variables but can be phenomenologically identified through recurring patterns in the user–technology relationship. The framework introduces three experiential categories:

1. *Immersion* (I) refers to the technological and perceptual conditions that allow users to be experientially embedded in an augmented or virtual environment. Following Slater and Wilbur (1997), immersion is closely tied to the role of technology in eliciting a sense of presence. In the context of ETs, immersion delineates the horizon of possible actions and perceptions within which interaction unfolds.

2. *Coherence* (C) concerns the perceived plausibility and internal consistency of the experience. An interaction is coherent when the system's behavior aligns with the user's expectations and embodied habits, maintaining a stable logic across actions and responses (Skarbez, Smith, and Whitton 2021). While immersion defines the range of possible actions, coherence ensures the experience feels credible and meaningful to the user.
3. *Entanglement* (E) designates the depth and quality of reciprocal adaptation between human and machine. Entanglement emerges when user actions leave traces on the system and, conversely, when the system's behavior reshapes the user's orientation and expectations in a continuous loop of mutual adjustment. This dimension is central to the ETs framework, as it captures the relational dynamics through which shared orientation and meaning may arise.

Taken together, these three dimensions do not describe separate stages, but coupled conditions shaping the relational dynamics of an interaction. Immersion and coherence function as enabling constraints: richer action–perception horizons (I) and greater experiential plausibility (C) typically stabilize the emergence of entanglement (E), but do not guarantee it, and entanglement may also occur in fragile and local forms when one of these constraints is weak. Importantly, only entanglement that is sufficiently supported by immersion and coherence can become sense-bestowing, allowing relational configurations that resemble a We-relationship to arise. In this sense, ICE provides an operational bridge between phenomenological concepts—such as intersubjectivity, embodiment, and sense-bestowal—and concrete interaction scenarios.

2.2 Applying the ETs Framework: A Prototypical Case Study

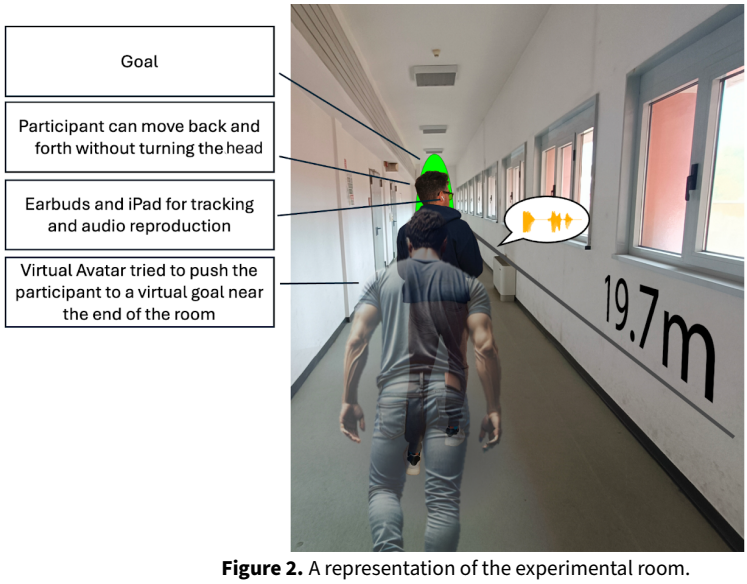

Figure 2. A representation of the experimental room.

To illustrate how the ICE framework operates in practice, we present a prototypical case study drawn from research on Sonic Interactions in Virtual Environments (SIVE) (Geronazzo

and Serafin 2023b). The case is not intended as a representative or statistically generalizable experiment, but as an exploratory investigation aimed at phenomenologically elucidating the conditions under which human–machine entanglement may emerge from the persuasive power of pervasive digital technologies of voice agents. The case study becomes particularly relevant in light of the relational capabilities of contemporary Large Language Models (LLMs), whose ability to sustain context-aware dialogue, simulate intentional stances, and imitate human communicative behavior significantly amplifies their potential to enact relational and empathic dynamics (Chang et al. 2026).

The study, as conducted in our laboratory (as in Figure 2), is described and followed by a discussion of its results through the proposed framework. The study is part of a broader Mixed-Method Research (MMR) design (Johnson, Onwuegbuzie, and Turner 2007), combining first-person accounts collected through post-interaction interviews with behavioral interaction data recorded during the session and analyzed within an HCI framework. However, the present paper focuses deliberately on the qualitative and phenomenological articulation of the case. The detailed MMR design, including experimental conditions, quantitative analyses, and statistical results, is developed in a separate study and will not be reproduced here.

The experiment involved a limited number of participants (N = 30), a choice that reflects the prototypical nature of the investigation rather than a limitation of scope. The aim was not to achieve statistical validation, but to provide a thick, experience-oriented description of human–machine interaction dynamics. From a phenomenological perspective, such prototypical cases are particularly valuable, as they allow the analysis to focus on structural features of experience rather than on population-level generalization.

This study adapts Kastanis and Slater's (2012) proxemics paradigm (Hall et al. 1968) from VR to an audio-based AR setting. The experience takes place in an audio-based mixed reality environment, where users interact with a virtual avatar through spatialized sound. Participants are invited to move freely within a physical corridor while wearing audio AR devices.[1] They are informed that another entity is present in the environment and instructed to engage with it until they feel comfortable. Crucially, users are not informed of the avatar's underlying objective, which consists in subtly orienting them toward a specific location within the space. Rather than being imposed, this orientation emerges progressively through interaction, as a shared directional alignment between user and system. Such an entangled relation is contingent upon sufficient levels of immersion and experiential coherence, which function as necessary conditions for meaningful engagement.

From the user's perspective, the interaction unfolds as a gradual process of orientation and mutual adjustment. The avatar emits spatial audio cues—such as footsteps, vocal calls, and breathing sounds—that invite movement and response. Although the system operates with a predefined goal, users typically experience the interaction not as coercive, but, under certain experiential conditions, as guiding and meaningful.

The case is particularly suitable for the ETs framework because it foregrounds relational dynamics rather than task completion. What matters is not whether the target point is

1. Using the Unity AR Foundation (https://docs.unity3d.com/Packages/com.unity.xr.arfoundation@6.0/manual/index.html) and SDN library https://github.com/msmhrtf/binsdn, the virtual space was precisely aligned with the physical room, enabling participants to navigate the augmented environment via an iPad and AirPods Pro, which also tracked head position and orientation. Although the virtual environment itself was not visible, participants could perceive and interact with it through spatialized audio cues which were customized in terms of room and listener acoustics.

reached, but *how* the interaction unfolds: whether the user experiences the avatar as a mere signal source or as a relational presence capable of shaping orientation and expectation. Immersion and coherence play a decisive role here. Without sufficient immersion, the avatar remains a disembodied sound; without coherence, its behavior appears arbitrary or intrusive. Only when both conditions are met does entanglement become possible.

2.3 ICE Traces and Phenomenological Interpretation

The phenomenological analysis of the experience can be articulated through ICE traces emerging from qualitative observations and post-interaction interviews. For analytical purposes, we describe four recurrent modes of adaptation as interpretive heuristics for characterizing patterns of interaction between user and system, rather than as fixed or exhaustive categories.

A) With respect to *immersion*, participants reported varying degrees of presence depending on technological comfort and sound localization. Even when technical imperfections were noted, many users described a tangible sense of being accompanied or followed, suggesting that immersion does not depend solely on technical perfection but on the overall integration of sensory cues.

B) *Coherence*, understood as the internal consistency of the avatar's behavior in relation to user expectations, emerged as a critical factor in shaping trust and engagement. When the avatar's behavior respected proxemic expectations and displayed temporal consistency, users were more likely to attribute intentionality and meaning to the interaction. Conversely, incoherent elements—such as an overly mechanical voice or ambiguous sound cues—risked disrupting the experiential flow.

C) The most significant dimension concerns *entanglement*. The analysis revealed four modes of adaptation: machine-to-user ($M \rightarrow U$), user-to-machine ($M \leftarrow U$), reciprocal ($M \leftrightarrow U$), and no adaptation. Entanglement traces were most evident when mutual adaptation occurred: users adjusted their movements in response to the avatar, while the avatar's behavior dynamically responded to the user's position and pace. In such cases, participants often described a sense of being "guided" rather than controlled, indicating the emergence of a shared orientation without explicit agreement. This relational configuration closely approximates the We-relationship described in the phenomenological framework. One-way adaptations, particularly $M \rightarrow U$, where the machine appears to follow the user's movements, can still open the possibility for reciprocity, especially when the user becomes receptive to the avatar's presence. Conversely, $M \leftarrow U$ reflects a less desirable asymmetry though unilateral projection. While a form of entanglement is still present, it is not sustained by adequate levels of immersion and coherence. The resulting relation is therefore characterized less by reciprocal co-regulation than by a projective investment on the part of the user.

D) Additionally, participants' affective responses played a crucial mediating role across all three dimensions. Participants reported a range of emotions, from curiosity and reassurance to discomfort or unease. Importantly, even negative affects did not necessarily disrupt entanglement; in some cases, they intensified engagement by heightening attention and responsiveness. This suggests that affectivity acts as a transversal modulator of immersion, coherence, and entanglement, influencing how relational meaning is constituted.

From a phenomenological perspective, the value of this case lies not in demonstrating optimal performance, but in revealing the conditions under which relational configurations between human and machine succeed or fail. Reaching the target point in the room, while important, is not sufficient: the "how" of the interaction—its relational modality—is crucial. For the We-relationship to emerge, the human and the machine must neither remain entirely separate nor merge into a single subjectivity, but instead recognize and sustain a common orientation that emerges within interaction, where human subjectivity is extended through relational forms of digital agency.

The experience shows that entanglement does not arise automatically from technical sophistication. Rather, it depends on the unfolding of interaction itself—on how users and systems reciprocally adjust within a shared experiential field. These findings support the view of ETs as relational configurations rather than autonomous agents or instrumental tools. The ICE framework thus functions as a descriptive lens through which the phenomenological structures identified in Section 1 can be traced within concrete interaction scenarios. This operational articulation prepares the ground for the ethical analysis developed in the next section.

3. Ethical Scenario: Agency, Responsibility, and Human-Centered Technology

The development of the ETs framework introduces not only novel technological and phenomenological perspectives, but also a set of ethical challenges that require careful and systematic consideration. These challenges go beyond the standard concerns associated with digital technologies and call for a rethinking of agency, in contexts where interaction is no longer merely instrumental but relational. Unlike traditional DTs, which primarily support decision-making through representational models, ETs operate within experiential and interactive domains where human perception, affectivity, and orientation are directly involved. This shift has significant ethical implications: when digital systems participate in shaping experience and meaning, ethical responsibility cannot be confined to isolated agents or discrete technical components, but must be understood as emerging within relational configurations.

3.1 Extended Agency and Relational Responsibility

At the heart of the ETs framework lies the notion of *extended agency*, a concept that captures the expansion of human capacities through the dynamic partnership with a non-human agent. ETs do not replace human agency, nor do they instantiate autonomous artificial subjectivity. Rather, they support forms of agency that emerge within a We-subject configuration, as articulated in Sections 1 and 2. Agency, in this context, is neither fully human nor purely machine-driven, but relationally constituted through interaction. A critical question arises here: can the ETs fully satisfy the philosophical definitions of agency? According to Latour, within his relational and social ontology, agency is defined as the capacity to make a difference in the course of action of some other agent (Latour 2005, p.71). From this standpoint, the machine component of the ETs can be recognized as an agent, capable of influencing the behavior of the human users within the relational network. Latour's criterion is intentionally minimal and relational. Phenomenological approaches, by contrast, propose

a stricter criterion: agency must be accompanied by a *sense of agency*, a form of self-awareness or experiential intentionality (Gallagher and Zahavi 2020, ch.8).

Can ETs meet this phenomenological standard? While the machine element of the ETs may lack full-fledged consciousness or awareness, their adaptive and responsive behavior can nonetheless contribute to experiential configurations in which users perceive a meaningful orientation shared with the system. This implies that human subjectivity is extended through forms of digital agency that remain structurally entangled with human intentionality. The human–machine relation thus evolves beyond mere instrumental interaction, creating conditions for *plural subjectivity* where agency is distributed but coherent. Therefore, while ETs may not meet the strictest definitions of conscious agency, they embody an operational sense of agency within the We-subject framework.

By contrast, the DT model operates predominantly through data replication and modeling, generating representations that influence the decisions of the users indirectly. DTs fit Latour's broader definition of agency, given their capacity to reshape practices and outcomes through data, but they fall short of the phenomenological criterion. There is no genuine sense of agency on the side of the DT, nor is there a direct, participatory relationship that reconfigures subjectivity. The interaction remains external and mediated.

The case study presented in Section 2 illustrates this point clearly. The emergence of reciprocal adaptation—rather than unilateral control—is crucial for the experience of being guided rather than constrained. Extended agency thus depends on immersion, coherence, and entanglement as experiential conditions, not on the attribution of mental states to the machine. Ethical evaluation must therefore focus on how the machine participates in shaping human action and experience.

This form of agency raises the question of how intentionality and decision-making responsibilities are distributed within this hybrid entity. As Merleau-Ponty (1968) and Sartre (1992) suggest, subjectivity is not static but emerges through interaction, perception, and engagement with others and the world. In the ETs framework, the We-subject embodies this phenomenological insight by allowing human users to navigate decisions with an augmented awareness of the environment and the alternatives at play. ETs enhance not only cognitive capacities but also the *situatedness* of agency, supporting decisions that are context-sensitive and relationally attuned.

This relational understanding of agency complicates traditional models of responsibility. The deeper integration between human and ETs increases the vulnerability of the user, particularly concerning autonomy and susceptibility to influence. One of the most distinctive features of the ETs agency is its *plural intentionality*, which emerges from both direct human–machine interaction and the indirect participation of designers, developers, and platform owners. This layered structure of agency complicates traditional notions of moral responsibility. Who is accountable when a decision made within a We-subject leads to unintended or harmful consequences?

Plural intentionality has the advantage of being more informed and shaped by a multifaceted set of influences: a broader range of variants and their nuances are taken into account. One might object that a greater plurality does not necessarily lead to better outcomes and that involving multiple agents in decision-making could complicate the attribution of individual responsibility (De Kerckhove 2021). We address this objection as follows. The contribution of external factors is not unique to the ETs; rather, it is an inherent characteristic of human

action itself, which is always influenced by socio-cultural environment, economic conditions, education, health status, profession, leisure activities, personal relationships, family circumstances, and more. Although the plurality of these factors does not deterministically define human action, it substantially contributes to the formulation of conscious and non-conscious decisions.

Accordingly, the ethical issue is not plurality as such, but the specific character of the influences at stake. In ETs, some influences are intentionally engineered, scalable, and potentially opaque, rather than socially sedimented or contingently acquired. This difference does not invalidate extended or plural agency, but it does require heightened ethical vigilance. Distributed agency must therefore be accompanied by traceability, transparency, and governance structures capable of accounting for how orientations, recommendations, and constraints are introduced into the interactional field.

Decisions and actions within ETs configurations are influenced not only by users, but also by system design, adaptive algorithms, and the intentions embedded by designers and developers. Responsibility cannot be reduced to a single locus; it must instead be distributed across the relational field in which action unfolds. However, distributed responsibility must not lead to diluted accountability. On the contrary, it requires clear governance structures, transparent design choices, and explicit acknowledgment of the roles played by different actors in shaping interaction.

3.2 Critiques of Digital Transformation and the ETs Response

Contemporary critiques of digital transformation often emphasize the risks posed by large-scale technological systems to human autonomy, social cohesion, and democratic life. Authors such as Fukuyama (2021) have highlighted the growing concentration of power in Big Tech companies and their capacity to influence political communication, while Han (2017, 2022) has stressed the alienating effects of digital platforms on subjectivity, social relations, and public discourse. These critiques are important, but they tend to frame digital technologies in predominantly apocalyptic terms, portraying them as external forces that inevitably erode human agency.

The ETs framework is not situated within this apocalyptic horizon. Rather than amplifying the logics criticized by Han or Fukuyama, ETs are explicitly designed to counter them at the level of interaction. They do not operate as opaque, large-scale infrastructures that extract data or manipulate behavior from a distance, but as localized, situated systems whose effects are directly perceptible within lived experience. Their ethical significance lies precisely in this scale and modality: ETs act within bounded interactional fields, where agency, influence, and responsiveness can be traced and critically assessed. Within this framework, digital technologies can be considered as facilitating personal growth, social connectedness, and novel forms of creativity (Botella et al. 2012).

Clarifying this point is particularly important in relation to current debates on *Artificial General Intelligence* (AGI). AGI is often presented as a trajectory toward autonomous, self-improving systems capable of open-ended reasoning and decision-making (Goertzel 2014; Kurzweil 2006). This trajectory raises significant ethical concerns, including the delegation of moral responsibility (Mahler 2022). ETs do not belong to this paradigm. They are not designed to approximate general intelligence, nor to function as independent loci of

reasoning or decision-making. On the contrary, ETs operate within the domain of narrow, task-specific AI, whose behavior is situated and relationally oriented.

This distinction is crucial for understanding both the ethical promise and the ethical risks of ETs. Recent discussions on LLMs illustrate what happens when this distinction is blurred. Although some studies have explored whether LLMs exhibit properties such as metacognition, situational awareness, or Theory of Mind (Chen et al. 2025; Strachan, Albergo, and Borghini 2024), further research has shown that behavioral complexity does not entail consciousness or inner experience (Li 2025). When users nonetheless interpret such systems as sentient or emotionally reciprocal, a misalignment emerges between what the system affords and what it appears to promise. Empirical studies on affective attachment to conversational agents, such as Replika,[2] have documented how this misalignment can lead to emotional distress and disillusionment when expectations of reciprocity are inevitably unmet (Ciriello et al. 2024).

ETs are explicitly designed to preempt the kind of relational misalignment observed in contemporary interactions with anthropomorphic AI systems. Rather than simulating human subjectivity or emotional reciprocity, they are structured around transparent relational boundaries and deliberate functional limitations. In this sense, ETs discourage the attribution of human subjectivity to machines and promote a form of *authentic interactivity* in which the system's constraints are experientially perceptible and acknowledged by the user. This design strategy is consistent with the ICE framework introduced in Section 2, which offers a phenomenologically grounded account of how ETs mediate experience beyond representational simulation. Rather than producing illusions of realism or human likeness, ETs modulate the sensory, cognitive, and affective conditions that support embodied co-regulation. Immersion and coherence establish the experiential ground for engagement, while entanglement emerges only insofar as interaction remains situated, intelligible, and relationally asymmetrical. Responsibility, within this framework, is not located in a single agent but arises from the relational field itself. Accordingly, ETs do not "possess" agency as an intrinsic property; they enact it within specific experiential configurations, resisting both anthropocentric projections and techno-centric reductions.

This architectural choice has concrete ethical and practical consequences. Rather than aspiring to autonomous decision-making or universal generality—as in AGI trajectories—ETs embrace situatedness, embodiment, and relational asymmetry. Unlike AGI-oriented trajectories, which aim at autonomous decision-making and generalized intelligence, ETs explicitly embrace situatedness, embodiment, and relational asymmetry. Their design draws on Sartre's notion of the We-subject (Sartre 1992), according to which shared intentionality does not require the fusion of subjects into a single consciousness, but rather the convergence of distinct perspectives within a common orientation. In the context of ETs, this means that human and artificial agents can participate in shared interactional structures without collapsing into a unified subjectivity or obscuring their ontological difference.

This distinction becomes particularly relevant when addressing what has been described as the "uncanny valley" of interactivity. Originally formulated in relation to humanoid robotics (Mori 1970), the uncanny valley has increasingly been observed in interactions with conversational agents and embodied AI systems, especially when human-like responsiveness is suggested without being sustainably supported (Wang, Shen, and Lee 2025; Zhang and

2. Replika is a generative AI chatbot designed to be a friend, partner, or mentor: <https://replika.com/>.

Pan 2025). In such cases, discomfort arises not merely from aesthetic resemblance, but from a deeper cognitive and affective dissonance: the system appears to invite forms of engagement—emotional reciprocity, mutual understanding, intentional responsiveness—that it cannot coherently sustain. The result is a breach in what might be described as the implicit relational contract between user and system. ETs are designed to avoid this breach not by intensifying anthropomorphic simulation, but by remaining deliberately on the near side of the uncanny valley. They enact responsiveness without implying consciousness, and guidance without simulating emotional reciprocity. This distinction between simulation and enactment is ethically crucial. Simulation presupposes a representational gap between appearance and function, whereas enactment treats interaction itself as the locus of meaning-making. Users are not encouraged to believe in a sentient other; instead, they are engaged in a process of mutual adjustment within clearly defined experiential constraints.

In this way, ETs avoid deceptive mimicry while enabling co-adaptive interactional forms that are both affective and embodied. Their responsiveness is grounded in lived temporality, spatial proximity, and perceptual attunement, reinforcing an ethics of presence rather than abstraction. The empirical relevance of authentic interactivity is supported by the prototypical study discussed in Section 2, where ETs did not aim to generate illusions of realism but to facilitate bodily and affective engagement through immersion, coherence, and entanglement. Through controlled feedback loops and proxemic cues, the system supported co-regulation without fostering affective dependency.

From a theoretical standpoint, this design stance also addresses persistent confusions surrounding intelligence and consciousness in AI systems. As Li (2025) cautions, behavioral or linguistic complexity—no matter how sophisticated—must not be conflated with inner experience. Failing to recognize this distinction risks reifying computational systems as sentient agents, with consequences ranging from emotional manipulation to the erosion of user autonomy. By explicitly rejecting both the illusion of artificial consciousness and the instrumentalization of personhood, ETs articulate a third path: one in which technological development is oriented toward relational clarity, ethical accountability, and shared presence rather than toward superintelligence or deceptive human likeness.

3.3 Industry 5.0 and the Human-Centered Turn

Insights from ethical debates surrounding DTs—particularly in healthcare—are instructive for understanding the ethical risks associated with ETs. Studies have highlighted issues such as data hypercollection, biased datasets, opaque algorithms, epistemic injustice, and the erosion of professional responsibility (Huang, Kim, and Schermer 2022; Popa et al. 2021). These insights underscore the need for robust data governance, algorithmic transparency, and patient-centered care models.

Although ETs are not limited to healthcare contexts, similar concerns arise wherever personalization, adaptive behavior, and user involvement are central. The entangled nature of ETs interactions can amplify both benefits and risks. On the one hand, adaptive systems may support more context-sensitive and meaningful experiences. On the other, they may obscure the origins of decisions or reinforce existing biases if transparency and fairness are not actively addressed. As Dignum (2019) emphasizes, explainability, accountability, and fairness must be treated as foundational design principles, especially in systems that shape user behavior and perception. From a phenomenological standpoint, epistemic justice (Fricker

2007) acquires particular importance. Users' lived experiences and interpretations must not be subordinated to algorithmic outputs or developer assumptions. Ethical ETs design requires attentiveness to how users make sense of interaction, how trust is established or undermined, and how vulnerability may arise through affective engagement.

The transition from Industry 4.0 to Industry 5.0 encapsulates the broader shift toward human-centered technologies. DTs are typically regarded as emblematic of the Fourth Industrial Revolution, itself an extension of the Third Industrial Revolution, characterized by the integration of the Internet of Things (IoT), cognitive computing, cloud computing, and AI. While DTs represent a key instrument in the progression of Industry 4.0 (Pires et al. 2019), there is a growing need to rethink this model in light of Industry 5.0, as recently emphasized by the European Commission.[3]

Unlike Industry 4.0, which remains largely machine-centered, Industry 5.0 aims for a *human-centered paradigm*, promoting an Internet of People (IoP) rather than a mere IoT infrastructure, as (Miranda et al. 2015) suggested years ago. Although research on human-centered technologies aimed at enhancing human well-being and restoring control over machines is still in its early stages, it has become a growing concern among researchers—even if it has not yet fully penetrated industrial practice (Alves, Lima, and Gaspar 2023). The transition from DTs to ETs mirrors this industrial shift: while DTs represent a crucial step towards a human-centered approach, they remain insufficient to fully realize it. In contrast, ETs explicitly aim to overcome the limitations of DTs and, much like Industry 5.0, are inherently designed within a human-centered framework.

Integrating the human sciences—such as philosophy—into the design and evaluation of ETs is essential for realizing this vision. The ethical landscape of ETs ultimately calls for an *ethics of entanglement*—a framework that recognizes the co-constitutive nature of human–machine relations. Such an ethics rejects both technophilia and technophobia, emphasizing instead reflective engagement with the affordances and limits of digital systems.

Key ethical commitments for ETs design include:

- Transparency and explainability: users should understand how interactional outcomes emerge within the We-subject.
- Accountability structures: clear mechanisms must exist to trace responsibility across human and non-human actors.
- Sustainability and equity: the development and deployment of ETs should prioritize environmental sustainability and social justice.
- Reciprocity and co-adaptation: ETs should be designed to adapt meaningfully to users.
- Emotional and affective attunement: ETs should respect and support the emotional dimensions of human experience.

In this sense, ETs do not propose a future of autonomous machines, nor a return to purely instrumental tools. They invite a reconfiguration of technological agency grounded in presence, interaction, and shared responsibility. By situating ethics at the level of experience rather than abstraction, the ETs framework contributes to a human-centered vision of digital technology that is both philosophically grounded and practically relevant.

3. See: <https://research-and-innovation.ec.europa.eu/research-area/industrial-research-and-innovation/industry-50_en>.

4. Conclusion

The ETs framework invites us to rethink the ontological and ethical stakes of human–machine relations in the age of digital transformation. In contrast to the logic of mirroring that characterizes the DT model, ETs inaugurate a relational paradigm wherein agency, subjectivity, and decision-making are no longer understood as the exclusive prerogative of the human subject, but as relational achievements co-constituted within a We-subjectivity that traverses the boundaries between the organic and the artificial.

This shift, as our phenomenological analysis suggests, is not merely technical but existential. The plural intentionality that emerges in the ETs dynamics is the mark of a subjectivity that is always already relational, attuned to an environment that is both mediated and transformative. In this sense, ETs do not simulate human capacities; rather, they participate in the modulation of human experience, configuring new pathways of perception, action, and meaning. The ICE triad—immersion, coherence, entanglement—crystallizes the conditions under which such hybrid subjectivity can flourish, opening a space where human and machine not only interact but intertwine their agencies.

From an ethical standpoint, the ETs propose a decisive alternative to the phantasm of AGI and its Promethean ambitions of autonomous superintelligence. The ETs refuse the model of mastery—whether human over machine, or machine over human—in favor of an ethics of entanglement, where responsibility is distributed, agency is shared, and reciprocity remains the guiding principle. In the face of the alienation and uncanny relationality often associated with AGI, ETs cultivate a space of intersubjective resonance, where the machine is not positioned as an absolute Other but as a participant of a We-subject in the unfolding of the world.

This reconfiguration aligns with the broader aspirations of Industry 5.0, seeking to overcome the reductionism of machine-centered paradigms in favor of human-centered and socially attuned technologies. In this sense, ETs should be understood as a conceptual framework capable of informing Industry 5.0's human-centered aspirations. Yet, as we have argued, this human-centeredness must itself be rethought—not as a reaffirmation of human exceptionalism, but as a co-constitutive horizon, where the human is always already entangled with the non-human.

In this light, the ETs do not merely extend the capacities of the human; they transform the very conditions of agency, foregrounding the relational matrices from which our actions, decisions, and responsibilities emerge. They challenge us to envision a technological future not dominated by control, prediction, or replication, but animated by co-becoming, ethical accountability, and shared world-making. In an era marked by the rapid evolution of mixed environments, ETs remind us that the future of technology cannot be separated from the future of subjectivity itself and that this future must be shaped, above all, by an ethics of entangled relations.

References

Abburu, Sailesh, Arne J. Berre, Michael Jacoby, Dumitru Roman, Ljiljana Stojanovic, and Nenad Stojanovic. 2020. COGNITWIN – Hybrid and Cognitive Digital Twins for the Process Industry. In *2020 IEEE International Conference on Engineering, Technology and Innovation (ICE/ITMC),* 1–8. Cardiff, UK: IEEE. https://doi.org/10.1109/ICE/ITMC49519.2020.9198403.

Alves, Joel, Tânia M. Lima, and Pedro D. Gaspar. 2023. Is Industry 5.0 a Human-Centred Approach? A Systematic Review. *Processes* 11 (1): 193. https://doi.org/10.3390/pr11010193.

Barad, Karen. 2007. *Meeting the universe halfway: Quantum physics and the entanglement of matter and meaning.* Durham: Duke University Press.

Barricelli, Barbara Rita, Elena Casiraghi, and Daniela Fogli. 2019. A Survey on Digital Twin: Definitions, Characteristics, Applications, and Design Implications. *IEEE Access* 7:167653–167671. https://doi.org/10.1109/ACCESS.2019.2953499.

Botella, Cristina, Giuseppe Riva, Andrea Gaggioli, Brenda K. Wiederhold, Mariano Alcaniz, and Rosa M. Baños. 2012. The Present and Future of Positive Technologies. *Cyberpsychology, Behavior, and Social Networking* 15 (2): 78–84. https://doi.org/10.1089/cyber.2011.0140.

Chang, Zhuang, Dominik O. W. Hirschberg, Kunal Gupta, Mehak Sharma, Kangsoo Kim, Huidong Bai, Li Shao, and Mark Billinghurst. 2026. Enhancing Perceived Empathy in Empathic Mixed Reality Agents via Context-Aware Adaptation. *IEEE Transactions on Visualization and Computer Graphics* 32 (2): 1569–1581. https://doi.org/10.1109/TVCG.2025.3646601.

Chen, Sirui, Shuqin Ma, Shu Yu, Hanwang Zhang, Shengjie Zhao, and Chaochao Lu. 2025. *Exploring Consciousness in LLMs: A Systematic Survey of Theories, Implementations, and Frontier Risks.* Version Number: 1. https://arxiv.org/abs/2505.19806 accessed February 23, 2026.

Ciriello, Raffaele, V. Suter, S. Hof, and G. Schwabe. 2024. Ethical tensions in human–AI companionship: A dialectical inquiry into Replika. In *Proceedings of the 57th Hawaii International Conference on System Sciences (HICSS),* 7333–7342.

Clark, Andy, and David Chalmers. 1998. The Extended Mind. *Analysis* 58 (1): 7–19.

De Kerckhove, Derrick. 2021. The personal digital twin, ethical considerations. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 379 (2207): 20200367. https://doi.org/10.1098/rsta.2020.0367.

Degli Innocenti, Edoardo, Michele Geronazzo, Diego Vescovi, Rolf Nordahl, Stefania Serafin, Luca Andrea Ludovico, and Federico Avanzini. 2019. Mobile virtual reality for musical genre learning in primary education. *Computers & Education* 139:102–117. https://doi.org/10.1016/j.compedu.2019.04.010.

Dignum, Virginia. 2019. *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way.* Cham: Springer.

Floridi, Luciano. 2010. *Information: A Very Short Introduction.* Oxford: Oxford University Press.

Frauenberger, Christopher. 2019. Entanglement HCI The Next Wave? *ACM Transactions on Computer-Human Interaction* 27 (1): 1–27. https://doi.org/10.1145/3364998.

Fricker, Miranda. 2007. *Epistemic Injustice: Power and the Ethics of Knowing.* Oxford: Oxford University Press.

Fukuyama, Francis. 2021. Making the Internet Safe for Democracy. *Journal of Democracy* 32 (2): 37–44.

Gallagher, Shaun, and Dan Zahavi. 2020. *The Phenomenological Mind.* London-New York: Routledge.

Geronazzo, Michele. 2022. Egocentric Audio in the Digital Twin of Virtual Environments. In *2022 IEEE 2nd International Conference on Intelligent Reality (ICIR),* 7–10. Piscataway, NJ, USA: IEEE. https://doi.org/10.1109/ICIR55739.2022.00017.

Geronazzo, Michele, and Stefania Serafin, eds. 2023a. *Sonic Interactions in Virtual Environments.* Human–Computer Interaction Series. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-04021-4.

Geronazzo, Michele, and Stefania Serafin. 2023b. Sonic Interactions in Virtual Environments: The Egocentric Audio Perspective of the Digital Twin. In *Sonic Interactions in Virtual Environments,* edited by Michele Geronazzo and Stefania Serafin, 3–45. Series Title: Human–Computer Interaction Series. Cham: Springer. https://doi.org/10.1007/978-3-031-04021-4_1.

Glaessgen, Edward, and David Stargel. 2012. The Digital Twin Paradigm for Future NASA and U.S. Air Force Vehicles. In *53rd AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference.* Honolulu, Hawaii: American Institute of Aeronautics / Astronautics. https://doi.org/10.2514/6.2012-1818.

Goertzel, Ben. 2014. Artificial General Intelligence: Concept, State of the Art, and Future Prospects. *Journal of Artificial General Intelligence* 5 (1): 1–48. https://doi.org/10.2478/jagi-2014-0001.

Hall, Edward T., Ray L. Birdwhistell, Bernhard Bock, Paul Bohannan, A. Richard Diebold, Marshall Durbin, Munro S. Edmonson, et al. 1968. Proxemics [and Comments and Replies]. *Current Anthropology* 9 (2/3): 83–108. https://doi.org/10.1086/200975.

Han, Byung-Chul. 2017. *In the Swarm: Digital Prospects.* Translated by E. Butler. Cambridge: MIT Press.

Han, Byung-Chul. 2022. *Infocracy: Digitalization and the Crisis of Democracy.* Translated by D. Steuer. Cambridge: Polity.

Hayles, Katherine. 1999. *How We Became Posthuman: Virtual Bodies in Cybernetics, Nature, and Informatics.* Chicago: University of Chicago Press.

Huang, Pei-hua, Ki-hun Kim, and Maartje Schermer. 2022. Ethical Issues of Digital Twins for Personalized Health Care Service: Preliminary Mapping Study. *Journal of Medical Internet Research* 24 (1): e33081. https://doi.org/10.2196/33081.

Husserl, Edmund. 1960. *Cartesian Meditations: An Introduction to Phenomenology.* Translated by D. Cairns. The Hague: M. Nijhoff.

Husserl, Edmund. 1983. *Ideas Pertaining to a Pure Phenomenology and to a Phenomenological Philosophy, First Book: General Introduction to a Pure Phenomenology.* Translated by F. Kersten. Dordrecht: Kluwer.

Jerald, Jason. 2016. *The VR book: human-centered design for virtual reality.* ACM books 8. New York: Association for computing machinery Morgan & Claypool publishers.

Johnson, R. Burke, Anthony J. Onwuegbuzie, and Lisa A. Turner. 2007. Toward a Definition of Mixed Methods Research. *Journal of Mixed Methods Research* 1 (2): 112–133. https://doi.org/10.1177/1558689806298224.

Kastanis, Iason, and Mel Slater. 2012. Reinforcement learning utilizes proxemics: An avatar learns to manipulate the position of people in immersive virtual reality. *ACM Transactions on Applied Perception* 9 (1): 1–15. https://doi.org/10.1145/2134203.2134206.

Kurzweil, Ray. 2006. *The singularity is near: When humans transcend biology.* New York: Penguin.

Latour, Bruno. 2005. *Reassembling the social: An introduction to actor-network-theory.* Oxford: Oxford University Press.

Li, Jingkai. 2025. Can "consciousness" be observed from large language model (LLM) internal states? Dissecting LLM representations obtained from Theory of Mind test with Integrated Information Theory and Span Representation analysis. *Natural Language Processing Journal* 12:100163. https://doi.org/10.1016/j.nlp.2025.100163.

Madni, Azad M., Carla C. Madni, and Scott D. Lucero. 2019. Leveraging Digital Twin Technology in Model-Based Systems Engineering. *Systems* 7 (1): 7. https://doi.org/10.3390/systems7010007.

Mahler, Tobias. 2022. Regulating Artificial General Intelligence (AGI). In *Law and Artificial Intelligence,* edited by B. Custers and E. Fosch-Villaronga, 521–540. The Hague: T.M.C. Asser Press.

Merleau-Ponty, Maurice. 1968. *The Visible and the Invisible.* Translated by A. Lingis. Evanston: Northwestern University Press.

Miranda, Javier, Niko Makitalo, Jose Garcia-Alonso, Javier Berrocal, Tommi Mikkonen, Carlos Canal, and Juan M. Murillo. 2015. From the Internet of Things to the Internet of People. *IEEE Internet Computing* 19 (2): 40–47. https://doi.org/10.1109/MIC.2015.24.

Mori, Masahiro. 1970. Bukimi no tani [The uncanny valley]. *Energy* 7:33–35.

Pires, Flávia, Ana Cachada, José Barbosa, António Paulo Moreira, and Paulo Leitão. 2019. Digital Twin in Industry 4.0: Technologies, Applications and Challenges, 721–726. Helsinki, Finland: IEEE. https://doi.org/10.1109/INDIN41052.2019.8972134.

Popa, Eugen Octav, M. van Hilten, O. Oosterkamp, and M.-J. Bogaardt. 2021. The use of digital twins in healthcare: socio-ethical benefits and socio-ethical risks. *Life Sciences, Society and Policy* 17 (1): 6. https://doi.org/https://doi:10.1186/s40504-021-00113-x.

Privitera, Alessandro, and Michele Geronazzo. 2024. Designing sonic interactions in intelligent reality with egocentric audio technologies. In *The Routledge Handbook of Sound Design,* 1st ed., edited by Michael Filimowicz, 219–251. London: Focal Press. https://doi.org/10.4324/9781003325567-16.

Sartre, Jean-Paul. 1992. *Being and Nothingness. A Phenomenological Essay on Ontology.* Translated by H.E. Barnes. New York: Washington Square Press.

Scheler, Max. 2008. *The Nature of Sympathy.* Translated by P. Heath. London-New York: Routledge.

Schutz, Alfred. 1967. *The Phenomenology of the Social World.* Translated by G. Walsh and F. Lehnert. Evanston: Northwestern University Press.

Skarbez, Richard, Missie Smith, and Mary C. Whitton. 2021. Revisiting Milgram and Kishino's Reality-Virtuality Continuum. *Frontiers in Virtual Reality* 2:647997. https://doi.org/10.3389/frvir.2021.647997.

Slater, Mel, and Sylvia Wilbur. 1997. A Framework for Immersive Virtual Environments (FIVE): Speculations on the Role of Presence in Virtual Environments. *Presence: Teleoperators and Virtual Environments* 6:603–616.

Stein, Edith. 1989. *On The Problem of Empathy.* Translated by W. Stein. Washington: ICS.

Strachan, James W.A., D. Albergo, and G. Borghini. 2024. Testing theory of mind in large language models and humans. *Nature Human Behaviour* 8:1285–1295. https://doi.org/10.1038/s41562-024-01882-z.

Wang, Xian, Luyano Shen, and Lik-Hang Lee. 2025. A Systematic Review of XR-Enabled Remote Human-Robot Interaction Systems. *ACM Computer Survey* 57 (11): 1–37. https://doi.org/10.1145/3730574.

Zahavi, Dan. 2015. You, Me, and We: The Sharing of Emotional Experiences. *Journal of Consciousness Studies* 22 (1-2): 84–101.

Zhang, Yongxiang, and Weiwei Pan. 2025. A Scoping Review of Embodied Conversational Agents in Education: Trends and Innovations from 2014 to 2024. *Interactive Learning Environments* 33 (7): 1–22. https://doi.org/http://dx.doi.org/10.1080/10494820.2025.2468972.

Philosophical Problems in Science (2025), **79**, 41–48
(Zagadnienia Filozoficzne w Nauce)
doi:10.59203/zfn.79.724

ZFN Philosophical Problems in Science

ARTICLE

Consciousness, normativity, and intrinsic meaning in large language models

Roman Krzanowski

The Pontifical University of John Paul II in Kraków, Poland
Email: rmkrzan@gmail.com

Abstract

Recent advances in large language models (LLMs), together with renewed speculation about artificial consciousness, have intensified debates concerning meaning and semantic intentionality in artificial systems. This paper advances a restricted and conditional thesis: even if LLMs were to possess some form of consciousness, this would not suffice to establish intrinsic semantic intentionality. To clarify this claim, I introduce a hieroglyph-copying thought experiment showing that consciousness, even when accompanied by goal-directed or procedural intentionality, does not guarantee normatively assessable semantic content. Drawing on this result, I argue that contemporary LLM architectures—despite their ability to generate contextually appropriate linguistic outputs—operate through probabilistic optimization processes that do not by themselves ground epistemic commitment or truth-directed representation. The argument does not deny the sophistication of LLM representations nor attempt to resolve competing theories of meaning. Rather, it isolates and defends an insufficiency thesis: consciousness alone does not ground intrinsic semantics. The apparent meaningfulness of LLM outputs is therefore best understood as dependent on human interpretive projection rather than as an internally grounded semantic achievement.

Keywords: large language models, intrinsic semantics, semantic intentionality, consciousness, normativity, philosophy of artificial intelligence

1. Introduction: The Question of Meaning in LLM Systems

Large language models (LLMs) (Vaswani, Shazeer, and Parmar 2017; Zhao et al. 2023; Chang et al. 2024; Fan et al. 2024) exhibit increasingly human-like linguistic capabilities. This development has reignited long-standing philosophical questions concerning the nature of understanding and meaning in artificial systems. Do these systems genuinely understand what they say, or do they merely execute highly sophisticated forms of pattern recognition without semantic comprehension?

Discussions of meaning in AI frequently invoke consciousness. If a system were conscious, the reasoning goes, it might also be capable of understanding or genuinely meaning its linguistic outputs. Some authors (Chalmers 2023; Hinton 2025) have suggested that advanced

AI systems may be exhibiting forms of proto-consciousness, raising the question of whether this could ground semantic understanding.

Throughout this paper, a distinction is drawn between procedural or goal-directed intentionality and semantic intentionality understood as contentful, truth-directed aboutness; the central claim concerns the insufficiency of the former for grounding the latter. By goal-oriented or procedural intentionality, I mean the directedness of actions toward outcomes or goals, which can be captured instrumentally or teleologically without presupposing intrinsic semantic content (Brentano 1874; Millikan 1984; Dennett 1987).

This paper argues that even if LLM systems were granted some form of consciousness, this alone would not be sufficient to secure intrinsic semantic meaning or genuine understanding. Intrinsic semantic meaning is meaning that belongs to a system itself, grounded in its own epistemic and normative relations, rather than imposed by an external interpreter. The paper does not attempt to settle the question of whether AI systems can be conscious. Rather, it examines whether consciousness—*if granted*—would be sufficient for intrinsic semantics or intentionality.

Although the paper refers to the Chinese Room argument, it does not seek to revive it, but rather to clarify a distinct confusion concerning the relation between consciousness and semantic competence.

To address this question, I first introduce the conditions under which a system can be said to possess meaning and understanding, emphasizing intrinsic semantic content, epistemic commitment, and normativity. I then introduce the Hieroglyph-Copying Scenario, which illustrates that consciousness and certain forms of intentionality may be present without semantic understanding. Next I examine large language models from an operational perspective and ask whether, given these conditions, they can be said to possess meaning, understanding, intentionality, or consciousness. For the sake of argument, I provisionally bracket the possibility that LLMs or future AI systems might exhibit some form of consciousness or intentionality. Finally, I argue that even if LLM systems—or future artificial systems—were to possess consciousness and intentionality, this would not by itself be sufficient to grant them intrinsic semantic meaning or genuine understanding.

The present argument is deliberately limited in scope. It does not attempt to refute all functionalist or informational theories of content, nor to settle the general metaphysics of intentionality. Its claim is conditional and restricted: even if large language models were granted forms of consciousness or goal-directed intentionality, this would not by itself suffice to ground intrinsic semantic content.

2. Meaning, Understanding, and the Conditions of Semantic Content

Discussions of meaning in artificial intelligence often proceed too quickly from observable linguistic performance to claims about understanding or semantic competence. In order to assess whether artificial systems can genuinely possess meaning, it is therefore necessary to clarify what is meant by *meaning* in the relevant philosophical sense and to distinguish it from merely functional or behavioral criteria.

The distinction drawn here parallels established contrasts in the literature between instrumental or derived intentionality (Dennett 1987), teleofunctional accounts (Millikan 1984), and truth-conditional or contentful representation (Dretske 1981; Fodor 1990). The present

use of "procedural intentionality" corresponds to forms of directedness that can be captured without invoking intrinsic truth-directed content, while "semantic intentionality" refers to states that are normatively assessable as correct or incorrect for the system itself.

In philosophy, meaning is not exhausted by the production of correct or useful outputs. To say that an expression is meaningful is to say that it is *about* something – that it represents, refers to, or purports to describe some object, concept, or state of affairs. This notion of aboutness is inseparable from normativity: meaningful representations can be correct or incorrect, true or false, appropriate or inappropriate. Meaning, in this sense, is not merely descriptive but evaluative, since it involves standards against which representations can succeed or fail.

Understanding is closely related to meaning but should not be identified with it. To understand is not simply to produce or manipulate meaningful symbols, but to stand in an epistemic relation to what those symbols represent. Understanding involves commitment to content, sensitivity to error, and responsiveness to reasons. A system that understands does not merely generate representations; it is answerable to how things are and can, in principle, recognize its own mistakes.

Intentionality is often invoked as the bridge between meaning and understanding, but here a crucial distinction must be drawn. On the one hand, there is procedural or goal-directed intentionality, understood as the directedness of actions toward outcomes or the systematic following of rules. On the other hand, there is semantic intentionality, which involves contentful mental states such as beliefs and judgments that are subject to standards of truth and justification. While semantic intentionality presupposes meaning, procedural intentionality does not. A system may act intentionally in the procedural sense—by following rules or pursuing goals—without thereby possessing semantic content or understanding.

Consciousness is frequently introduced into these debates as a potential grounding for meaning. It is often suggested that if a system were conscious, it might thereby be capable of genuine understanding. However, consciousness alone does not determine whether a system possesses intrinsic semantic content. Even if a system were granted some form of conscious experience, it would still be an open question whether its representations are normatively constrained – whether they function as beliefs or judgments that can be correct or incorrect for the system itself. Consciousness may be relevant to meaning, but it is not sufficient for it.

In human agents, semantic meaning is typically embedded within broader epistemic practices: interacting with the world, forming beliefs, correcting errors, and taking responsibility for representations. These practices illustrate how meaning is ordinarily grounded, but they do not by themselves establish necessary conditions that any system must satisfy in order to possess meaning. What matters is not the particular biological or social route by which meaning is achieved, but the presence of epistemic commitment and normativity.

The upshot is that neither linguistic performance, nor procedural intentionality, nor even consciousness guarantees intrinsic semantic meaning. A system may produce symbols that are meaningful to others, act intentionally with respect to its tasks, and even enjoy conscious states, while still lacking understanding. The following section introduces a thought experiment—the Hieroglyph-Copying Scenario—that makes this dissociation explicit and clarifies why meaning cannot be reduced to consciousness or intentional action alone.

3. The Hieroglyph-Copying Scenario

The Hieroglyph-Copying Scenario unfolds as follows. A person consciously copies Egyptian hieroglyphs without knowing their meaning. Although these signs may carry meaning for others (for example, Egyptologists), they remain meaningless to the person who copies them. The copier is conscious and intentionally draws the hieroglyphs, yet they do not convey any meaning *for him or her*. The person is not communicating a meaningful message, nor does he or she understand what the symbols stand for.

Of course, for someone versed in Egyptian hieroglyphs, the copied inscriptions may have semantic content. However, this meaning is not possessed by the copier. Rather, it is imposed by an external interpreter. The meaning of the hieroglyphs, in this case, belongs to the observer, not to the agent who produces them.

What do we learn from this scenario? We learn that even conscious agents can generate symbols that are meaningful *for others* without understanding them themselves. Moreover, systems that are conscious and intentional with respect to their actions do not thereby possess intrinsic semantic content. Consciousness and goal-directed intentionality are therefore not sufficient for intrinsic semantics.

The force of the scenario does not depend on introspective access as a necessary condition of meaning. The point is not that lack of meta-awareness entails absence of first-order content, but that consciousness alone—even when present—does not secure truth-directed, normatively assessable representation. The example isolates the insufficiency of consciousness, not the necessity of phenomenology for semantics.

4. Large Language Models: An Operational Overview

Large Language Models like GPT-4 or Claude (or software systems based on the similar architecture or any other systems with LLM architecture; see for a list of such systems: (Hadi et al. 2023; Ariffud 2026) are trained on vast corpora of human-generated text. They learn statistical associations between words, phrases, and contexts, allowing them to generate fluent and contextually appropriate responses (see the review of LLM systems by: Raiaan et al. 2024; Vaswani, Shazeer, and Parmar 2017).

LLMs are composed of large numbers of correlated parameters structured as multi-dimensional data arrays. Through training, these systems learn statistical associations between tokens (word fragments, words, or sequences of words), phrases, and sentences, enabling them to generate fluent and contextually appropriate responses over long sequences (Raiaan et al. 2024; Vaswani, Shazeer, and Parmar 2017). The process by which LLMs "learn" is sometimes likened to statistical pattern recognition (see e.g. Chang et al. 2024; Fan et al. 2024; Zhao et al. 2023; Hadi et al. 2023; Raiaan et al. 2024).

In transformer architectures (Vaswani, Shazeer, and Parmar 2017), tokens are mapped to high-dimensional embedding vectors, whose geometric relations encode statistical co-occurrence patterns. The attention mechanism dynamically weights contextual relations between tokens, enabling context-sensitive prediction of subsequent tokens. These operations allow for remarkable linguistic performance but remain governed by probabilistic optimization objectives rather than truth-evaluative commitments.

5. LLM and Understanding

LLM algorithms optimize predictive accuracy without any reference to the semantic content of the data. This description concerns their operational training objective, not a metaphysical denial of representational structure. Their "knowledge" is a byproduct of massive-scale correlation, not comprehension. The presence of human-in-the-loop reinforcement, such as RLHF (Reinforcement Learning from Human Feedback), may refine outputs, but it does not endow the system with intentionality. Consequently, while their outputs may appear coherent, even insightful, they are not meaningful in the sense defined by philosophical theories of language use.

The apparent meaning in LLM output arises from the interpretation imposed by human users who ascribe meaning to language based on context, experience, and shared conceptual frameworks. The LLM does not thereby know, believe, or mean anything in the intrinsic semantic sense defined above (see (Marcus and Davis 2020; Wolfram 2023). LLMs lack necessary properties requires for a system (any system) to possess meaning.

The LLM does not evaluate whether a response *makes sense* in a semantic or epistemic manner at any stage in this process. This is because it lacks any representation of meaning. Rather, the system optimizes numerical objectives defined over probability distributions within a multilayer computational procedure. The guiding criterion for output generation is, therefore, not semantic understanding. It is statistical optimization – typically the maximization of likelihood or related objective functions.

This point can be illustrated by a simple observation. Consider perturbations in the model's probability space (understood here as a high-dimensional vector space rather than a physical or mathematical "complex" space). These perturbations can yield outputs that are grammatically well-formed yet semantically incoherent. The truth of such an output is also evaluated with respect to the statistical coherence of syntax rather than meaning. This demonstrates that grammaticality and surface coherence are consequences of learned statistical regularities, not indicators of understanding.

6. Philosophical Thought Experiments: Ants, Chinese Rooms, and Hieroglyphs

These considerations do not introduce an independent objection to AI meaning, but reinforce the insufficiency result established by the Hieroglyph-Copying Scenario.

Two thought experiments illustrate the disconnect between symbol manipulation and genuine understanding:

1. Putnam's Ant (Putnam 1981): An ant wandering randomly on the sand may trace a shape resembling Winston Churchill. Despite the resemblance, the ant did not draw Churchill. The interpretation is imposed by an observer, not intended by the ant.
2. Searle's Chinese Room (Searle 1980; Cole 2024): A person inside a room manipulates Chinese symbols using a rulebook without understanding the language. From the outside, the person appears to understand Chinese, but inside, no understanding occurs. This demonstrates that syntactic manipulation does not entail semantic comprehension. Note that the Chinese Room argument does not equate to the claim that linguistic meaning, in a general sense, is solely internal to the mind, the brain, or a system. Searle has repeatedly emphasized the social and institutional dimensions of language, stressing its public, conventional, and culturally embedded character (e.g., Searle 1995). His broader

philosophy of language, therefore, supports a mixed internal–external account of meaning. The Chinese Room argument itself is more narrowly focused. It concerns systems that process linguistic inputs exclusively by means of internally sufficient formal procedures. These are procedures that are complete with respect to input–output mapping but lack any semantic or epistemic perspective. Any system whose operation is exhausted by such internally complete procedures (as described in the Chinese Room) cannot thereby generate intrinsic meaning or understanding.

These examples underscore a critical point: neither syntax, statistical semantics (what LLM systems create from input data), nor even consciousness alone ensures meaning. The Hieroglyph-Copying Scenario is particularly relevant as in this scenario a conscious person is responsible for generating "meaningful signs without understanding them" demonstrating that even conscious system may generate meaningful, for someone else, symbols without understanding their meaning. What is missing is the agent's semantic intentionality.

7. LLMs and understanding—Objections

A common objection to denying that LLMs possess meaning is that LLMs generate outputs by training on meaningful human language and thus inherit meaning through this process. However, this objection fails to grasp the distinction between statistical structure and semantic content. An LLM trained on syntactically correct but semantically meaningless data—such as a corpus of meaningless but grammatically plausible text—would still generate fluent responses. Even if such a corpus is hypothetical, the point illustrates the distinction between statistical regularity and semantic grounding. Yet these responses would not be meaningful, as the model cannot distinguish between meaningful and meaningless input without intentional or social context. Its performance is driven by probabilistic patterns, not by understanding or intention.

Another objection suggests that LLMs generate a kind of meaning through what is sometimes called *statistical semantics*. However, this term is misleading, as it bears little resemblance to *semantics* in the traditional philosophical or linguistic sense. Technically, statistical semantics refers to patterns of co-occurrence—statistical relationships between tokens, words, or phrases—not to meaning as understood through reference, intentionality, or comprehension. The atomic elements of LLMs—tokens and word fragments—are assembled based on probabilistic associations, not semantic understanding. Since semantics plays no role during the training or compilation of an LLM's dataset, it is a dubious claim to suggest that meaning somehow "emerges" from inherently non-semantic processes. Yet this is precisely what some proponents of LLMs—including certain developers and users—appear to argue, presenting it as an emerging capacity or ability of such systems (Schaeffer, Miranda, and Koyejo 2023; Wei et al. 2022).

One may argue (a functionalist) that normativity itself could emerge from sufficiently complex functional organization. The present argument does not deny this as a logical possibility. Rather, it maintains that contemporary LLM architectures, as currently understood, do not instantiate mechanisms that would ground epistemic commitment or truth-directed representation at the system level.

One might also object that intrinsic semantic intentionality could emerge from sufficiently complex computational systems. However, the argument presented here concerns a concep-

tual insufficiency rather than a limitation of scale or architecture: increases in complexity or statistical sophistication do not, by themselves, generate normatively binding epistemic commitments. Without an account of how such normativity could arise from purely formal operations, appeals to emergence remain explanatorily incomplete.

8. Conclusion: Consciousness Without Meaning

This paper set out to examine whether recent claims concerning consciousness and intentionality in large language models can ground genuine meaning or understanding in artificial systems. Rather than attempting to settle the question of whether LLMs are, or could become, conscious, the argument adopted a conditional strategy: even if consciousness and certain forms of intentionality were granted to artificial systems, intrinsic semantic meaning would not thereby follow.

The Hieroglyph-Copying Scenario was introduced to clarify this point. It demonstrated that a conscious agent may intentionally and systematically manipulate symbols that are meaningful to others without understanding them or possessing semantic content for itself. This dissociation shows that consciousness, even when combined with procedural or goal-directed intentionality, is insufficient for intrinsic semantics. Meaning requires more than the capacity to experience or to act intentionally; it requires epistemic commitment, normativity, and the ability to stand in a truth-directed relation to what is represented.

When applied to large language models, this result suggests a principled limitation. LLMs can generate linguistically appropriate and contextually sensitive outputs, but such performance does not by itself amount to understanding. Even if future systems were to exhibit forms of consciousness or intentional states, this would not automatically provide them with intrinsic semantic content. Without epistemic commitment—without beliefs that can be correct or incorrect for the system itself—semantic meaning remains externally imposed rather than internally grounded.

Takeaway message is that "Even if consciousness and certain forms of intentionality are granted, intrinsic semantic meaning does not follow."

The broader implication is that debates about meaning in AI should not focus exclusively on architectural complexity, scale, or even consciousness, but on the conditions under which semantic content can be intrinsic to a system. By separating consciousness from semantic competence, the paper clarifies a persistent conceptual confusion in discussions of artificial intelligence. Meaning without epistemic agency is only apparent meaning: a projection of human interpretation onto systems whose operations, however sophisticated, remain formally and normatively uncommitted.

Statement on the use of AI tools

Parts of the manuscript were linguistically revised with the assistance of AI-based language tools. All philosophical ideas, arguments, and conclusions are the author's own, and the author takes full responsibility for the content of the paper.

References

Ariffud, Muhammad. 2026. *What are large language models (LLMs), how do they work, and popular use cases.* https://www.hostinger.com/tutorials/large-language-models accessed February 22, 2026.

Brentano, Franz. 1874. *Psychology from an Empirical Standpoint.* Translated by A.C. Rancurello, D.B. Terrell, and L.L. McAlister. London: Routledge & Kegan Paul.

Chalmers, David. 2023. *Could a large language model be conscious?* https://arxiv.org/abs/2303.07103 accessed June 22, 2026.

Chang, Yupeng, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, et al. 2024. A Survey on Evaluation of Large Language Models. Version Number: 9, *ACM transactions on intelligent systems and technology* 15 (3): 1–45.

Cole, David. 2024. *The Chinese Room Argument.* https://plato.stanford.edu/entries/chinese-room/ accessed February 22, 2026.

Dennett, Daniel. 1987. *The Intentional Stance.* Cambridge: MIT Press.

Dretske, Fred I. 1981. *Knowledge and the Flow of Information.* Cambridge: MIT Press.

Fan, Lizhou, Lingyao Li, Zihui Ma, Sanggyu Lee, Huizi Yu, and Libby Hemphill. 2024. A Bibliometric Review of Large Language Models Research from 2017 to 2023. *ACM Transactions on Intelligent Systems and Technology* 15 (5): 1–25. https://doi.org/10.1145/3664930.

Fodor, Jerry A. 1990. *A Theory of Content and Other Essays.* Cambridge: MIT Press.

Hadi, Muhammad Usman, Qasem Al Tashi, Rizwan Qureshi, Abbas Shah, Amgad Muneer, Muhammad Irfan, Anas Zafar, et al. 2023. *Large Language Models: A Comprehensive Survey of its Applications, Challenges, Limitations, and Future Prospects.* https://doi.org/10.36227/techrxiv.23589741.v4.

Hinton, Geoffrey. 2025. *'Godfather of AI' predicts it will take over the world.* https://www.youtube.com/watch?v=vxkBE23zDmQ accessed February 22, 2026.

Marcus, Gary, and Ernest Davis. 2020. *GPT-3, Bloviator: OpenAI's language generator has no idea what it's talking about.* https://www.technologyreview.com/2020/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-ai-opinion/ accessed February 22, 2026.

Millikan, Ruth Garrett. 1984. *Language, Thought, and Other Biological Categories: New Foundations for Realism.* Cambridge: MIT Press.

Putnam, Hilary. 1981. *Reason, truth and history.* Cambridge: Cambridge University Press.

Raiaan, Mohaimenul Azam Khan, Md. Saddam Hossain Mukta, Kaniz Fatema, Nur Mohammad Fahad, Sadman Sakib, Most Marufatul Jannat Mim, Jubaer Ahmad, Mohammed Eunus Ali, and Sami Azam. 2024. A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges. *IEEE Access* 12:26839–26874. https://doi.org/10.1109/ACCESS.2024.3365742.

Schaeffer, Rylan, Brando Miranda, and Sanmi Koyejo. 2023. Are emergent abilities of large language models a mirage? *Advances in Neural Information Processing Systems* 36:55565–55581.

Searle, John. 1980. The intentionality of intention and action. *Cognitive Science* 4:47–70.

Searle, John. 1995. *The construction of social reality.* London: Penguin Books.

Vaswani, Ashish, Noam Shazeer, and Niki Parmar. 2017. Attention is all you need. In *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems,* edited by Ulrike Luxburg, Isabelle Gyuon, and Samy Bengio, 6000–6010. New York: Curran Associates Inc.

Wei, Jason, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, et al. 2022. *Emergent Abilities of Large Language Models.* Version Number: 2. https://arxiv.org/abs/2206.07682 accessed February 23, 2026.

Wolfram, Stephen. 2023. *All-In Summit: Stephen Wolfram on computation, AI, and the nature of the universe.* https://www.youtube.com/watch?v=%7B2%7DcQmQIYNI5M accessed February 22, 2026.

Zhao, Wayne Xin, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, et al. 2023. *A Survey of Large Language Models.* Version Number: 18. https://arxiv.org/abs/2303.18223 accessed February 23, 2026.

Philosophical Problems in Science (2025), **79**, 49–58
(Zagadnienia Filozoficzne w Nauce)
doi:10.59203/zfn.79.749

ARTICLE

Towards happiness: Ancient Greek values in the age of computational intelligence supporting happiness—selected issues

Teresa Marcinów

Wrocław University of Science and Technology, Poland
Email: teresa.marcinow@pwr.edu.pl

Abstract

This article offers a selective, concise comparison of the ancient Greeks' ideas about happiness with contemporary metrics and AI-mediated wellbeing apps. It argues that these schools' views of happiness provide criteria for assessing the trajectory of today's approaches of development to the subject of happiness. Using a critical–analytical approach, it presents and compares those schools with major happiness metrics (e.g., World Happiness Report, Better Life Index, Happy Planet Index, Gallup Global Emotions) and popular wellbeing applications (Headspace, Calm, Insight Timer, Happify). The findings clarify where ancient insights align or clash with current practices and offer implications for wellbeing technologies. The paper also evaluates whether AI can support eudaimonia without eroding autonomy or commodifying the inner life. The article raises awareness that we are entering a new phase of understanding happiness, in which technology influences our choices and becomes not only a tool, but also a co-creator of our everyday experiences.

Keywords: happiness, eudaimonia, hedonism, happiness metrics, wellbeing apps

Introduction

The search for happiness is one of the key themes of philosophical reflection, an enduring concern of everyday life. Although the philosophers of ancient Greece proposed elaborate and diverse visions of happiness, contemporary society aided by technology and artificial intelligence algorithms, is attempting to reinterpret for a new generation what happiness is and under what conditions it can be experienced.

In order to achieve these research objectives, several main areas have been identified. In the first part of the article, the researcher attempts to present the main schools of ancient philosophical thought on happiness. The author refers to representatives of hedonism, Aristotle's eudaimonia, and the Stoics, and analyzes their possible references in contemporary, popular applications and programs devoted to the subject of happiness. In the second part, an attempt is made to answer the following questions: Do contemporary reports and applications

devoted to happiness, made possible by artificial intelligence, combine the legacy of the ancient Greeks with data optimization? What does the currently proposed path to happiness look like, and what does the practice of caring for well-being and personal development look like according to popular tools of the digital age? The third section discusses selected reports and applications and evaluates their potential impact on human life.

This article is a contribution to further research; it does not claim to provide definitive conclusions but rather examines the transformation of the understanding of happiness in the context of artificial intelligence. Ancient models of happiness, which were embedded in a social structure that took into account some presumed objective good and assumed some natural ends of human life, have been transformed by the applied concept of happiness as well-being. Happiness is achieved by the optimization of highly individualized class of goods.

The legacy of antiquity also includes specific positions on the character of thought, such as wisdom, value, happiness, virtue, pleasure, and human nature. Each of these systems proposes a slightly different way of life, answering the question of the meaning and manner of living a good life (Haybron 2007; Swoboda 1997).

Four models of achieving happiness in ancient Greece will be demonstrated by presenting the search for answers to the question of what a good life is, as well as selected applications that have the broadest impact, offering support in achieving happiness and appearing as a proposal in response to a request for support in the search for happiness.

It is worth noting that the very concept of happiness has varied greatly over the centuries – the concept often covered radically different paths to a good life, combining both the desire for fulfillment and a sense of harmony with oneself. Happiness is also referred to as success, intense joy, the greatest amount of goods, and satisfaction with life (Tatarkiewicz 2005) but certain ambiguity almost always accompanies any attempt to describe term "happiness" (Mordarski 2023; White 2008; Karaś 2019; Krasucka 2014). The article concludes with a brief synthesis of findings.

The hedonism of Aristippus and Epicurus, eudaimonia, and stoicism

Ancient philosophy, from Aristippus of Cyrene, Aristotle, and Epicurus to the Stoics, developed various schools of thought on happiness. Among them, three currents of understanding happiness gained particular significance: hedonism shaped by Aristippus and Epicurus, eudaimonia, and stoicism, which was continued by Roman thinkers. The proposed assumptions remain an indispensable point of reference in analyzing the experience of happiness to this day.

The hedonistic school associated with Aristippus of Cyrene (c. 435–356 BC) understands happiness as the maximization of mainly physical pleasures and the avoidance of pain; unlike Epicureanism, the mere absence of pain is not considered a good in itself. The word "hēdonē" itself, Latin "voluptas", means "delight". The hedonistic approach has a tendency towards reductionism, perceiving man as part of the material world and not overly emphasizing the spiritual dimension that reaches beyond the perceptible, earthly horizon. Matter and temporality begin to determine the dimensions of happiness, and the senses are the path to its attainment. Pain and pleasure, effort and comfort stand in opposition and reveal the tension that accompanies humans in their pursuit of fulfillment. This necessary distinction refers to the contrast between good, understood as pleasure, and evil, understood as pain and

discomfort. Happiness as freedom from unpleasant experiences becomes the goal of existence. In the face of the transience and fleetingness of life, hedonists directed their attention towards the present – they found the most important meaning in the "here and now" as the only certain circumstance and available source of happiness.

According to the Cyrenaics, because life is full of unknowns and uncertainties, happiness should not be tied to the good of others or to the success of social causes (Tatarkiewicz 2005). On the other hand, cultivating virtues is justified insofar and only insofar as they can help in achieving pleasure. Nevertheless, it would also be wrong to accuse the Cyrenaics of completely losing themselves in pleasure. Based on preserved accounts, it can be assumed that they preached the maxim of Epictetus: "It is the nature of the wise to resist pleasures, but the foolish to be a slave to them".

Hedonism is a type of philosophical position that gained many adherents.[1] Among them was Epicurus (341–270 BC), who proposed a new, more moderate version of this idea. Like his predecessors, he referred to the concept of happiness as the absence of suffering, but his vision was far removed from sensual ecstasy and was not extreme in nature. In his account of happiness, he emphasized aponia and ataraxia, appreciating the profound value of interpersonal relationships, reflection, and peace. He can be credited with refining or even "civilizing" hedonism, elevating it from the level of sensual gratification and transforming it into a reflective system. It was no longer just the immediate pleasure associated with the present that gained significance, but also its mental or "spiritual" dimension. On this view it is possible to both refer to the past and look to the future. The means leading to a happy life rely not only on virtue, but also on the way we practice virtues in the polis, building a community.

A prudent approach to sensuality directs a person towards inner riches: "If you live according to your nature, you will never be poor; if you live according to human opinions, you will never be rich" (Seneca 1917).[2] In summary, Epicurus' proposal is freedom from the world's preoccupations, peace achieved through a moderate life in which happiness is manifested as freedom from suffering.

This type of reflection has found some place in psychology, and it can be said that, as it clearly favors pleasant mental states, it has shaped contemporary thinking about human well-being. Pleasant sensations and positive emotions have been recognized as conducive to mental health (Fredrickson 2002).

Another important philosophical concept concerning happiness is *eudaimonia*. According to Aristotle (384–322 BC), every action aims at some good; the highest human good is *eudaimonia*, often described as self-sufficient (*autarkeia*) (Arystoteles 1996). The relationship between the questions 'What is *eudaimonia*?' and 'What is the good for human beings?' frames more detailed inquiry.[3]

1. The legacy of hedonism found its place in Jeremy Bentham's work on utilitarianism as teaching the greatest pleasure for the greatest number of people and was also included in the philosophy of John Stuart Mill, who distinguished between higher and lower values – he classified intellectual and spiritual values as higher, and sensual values as lower.

2. Seneca clearly states that this is "a saying of Epicurus."

3. When referring to Aristotle's views, it should be remembered that *eudaimonia* is based on a modified version of Plato's ideas, with which it is in some dialogue (it also serves as a counterpoint to many later observations and concepts of happiness). Plato's thought forms the axiological foundation through its understanding of good as an

In the *Nicomachean Ethics*, Aristotle presents human action as a sequence of aspirations, each aiming at a particular end: "All actions and choices seem to be directed toward some good—and that is why it is rightly called good, that toward which everything is directed" (Arystoteles 1996, EN I.1, 1094a1–3).

The intermediate good that has already been achieved leads to the next good, and at the end of this chain of goals and goods is the highest good. The actions taken and the path traveled therefore have their ultimate goal, which is happiness: "Happiness turns out to be something perfect and self-sufficient, being the goal towards which everything else tends" (Arystoteles 1996, EN I.7, 1097b20–22). In this way, Aristotle rejects the idea of endless pursuit; human life has an internal direction, and happiness can be achieved through the perfection of virtues and rational action.

The Stagirite does not ground his guidelines for achieving happiness on actions based on pleasure and sensual enjoyment, because these actions could be attributed to animals, and humans, as rational beings, are obliged to use this unique characteristic. The acquisition of wealth cannot be considered as the goal of human activity either, since it serves only as a means and therefore has no ultimate value. The pursuit of honors is merely a confirmation by others of what one has already recognized as an achievement and is also rejected by Aristotle. What, then, is worth striving for? Aristotle identifies an activity distinctive of human beings – namely, rational activity in accordance with virtue.

According to the philosopher, a person who lives happily strives to develop and use his or her mental and moral potential for these are spiritual goods, the most highly valued ones. The development of a stable disposition to act in accordance with reason is a virtue and is understood as a kind of good. People are not born with these virtues; they become virtuous because they work on their inner resources and must constantly be vigilant so that some of them are not lost. According to Aristotle, virtues can be divided into two kinds: intellectual and ethical. The former, once acquired, remain with a person permanently, while the latter—e.g., justice, courage, generosity—require not only work to develop them, but also effort and care to maintain them in a permanent state. To summarize Aristotle's path to happiness, it is self-knowledge and self-improvement, coming to terms with one's weaknesses, recognizing and developing one's strengths, and following the path of the "golden mean", which functions as a measure for ethical virtues and a guide to their development.

The importance of *eudaimonia* has remained significant for subsequent generations, because it concerns not only the fleeting aspects of happiness but also the long-term cultivation of character and the discovery of one's potential. The Stoics offered a different way of thinking about happiness. The philosophical movement initiated by Zeno of Citium (c. 335–263 BC) was derived from many earlier schools. It drew on the materialistic understanding of reality (materialistic monism) and critically interacted with rival views. The Stoics considered rational beings to be the most perfect of all possible beings – and uniquely capable of moral development. This perfection is most fully expressed in the ability to distinguish between what is instinctive and animalistic and what is unique to man. Happiness, therefore, is not something that an animal can achieve, but something specifically human.

Although many actions are morally neutral (*adiaphora*), the path to happiness leads through the rational cultivation of virtue. Particularly valued were practical wisdom, moderation,

objective value, the positing of reason as the central authority in the structure of the soul, and the purposefulness of existence.

justice, and courage – which also play an important place in Plato's work. They proposed understanding nature as a rational ordered cosmos (*logos*).

Emotions were also considered natural, but capable of disrupting rational behavior and resulting in dependence on external circumstances. In response to this risk, the Stoics proposed concept of *apatheia* – independence from destructive passions. Everything that is not related to the pursuit of virtue, which is an end in itself and whose practice leads to the attainment of good and happiness, should be indifferent.

To summarize this part of the study, the views can be presented in a synthetic formula.

Table 1. Values leading to happiness according to selected ancient philosophers

Philosopher	Definition of happiness	Core values	Purpose of life
Aristippus	Pleasure as the highest good; sensual experience in the present moment	Intensity of experience, freedom, control over pleasures	Maximizing pleasure and avoiding pain
Aristotle	Fulfillment through moral and intellectual virtues	Reasonableness, moderation, justice, friendship	Full development of human potential
Epicurus	Pleasure understood as the absence of suffering (*ataraxia*) and peace of mind	Moderation, friendship, simplicity of life	Lasting peace and freedom from suffering
Zeno of Citium	Living in harmony with nature and reason; acceptance of fate	Virtue, discipline, resistance to emotions, acceptance of fate	Peace of mind and inner freedom

1. Happiness as a measurable phenomenon in the private and social spheres

Ideas from classical philosophy were partly taken up by psychology with the rise of the humanistic movement in the twentieth century. Psychology, which has developed as an empirical science and addressed topics such as cognitive processes, emotions, and sources of motivation. There was also an interest in happiness understood as an element of full human development.

One of the first concepts in this area was proposed by Abraham Maslow (1954), who formulated the hierarchy of needs and emphasized the importance of self-actualization. Work on measuring happiness continued throughout the twentieth century (Angner 2011), gathering momentum in the 1970s and accelerating further in the 1980s. During that period, the possibility and need to measure happiness arose, leading to classify individuals, groups, and ultimately societies as more or less happy. As a result, the already extensive research areas of this concept gained new content and dependencies between new topics appeared (Mitchell et al. 2013).

Initially, measurements were based mainly on questionnaires and interviews, but over time, neurobiological methods were also used to study happiness at the level of neuroactivity (Richter et al. 2024). The inclusion of computational methods made it possible to create statistical models and map relationships, extracting patterns of satisfaction from the private sphere, which soon became a helpful market tool. These approaches were quickly taken up across commercial domains, especially marketing and management.

In this way, happiness—more clearly than in past eras—ceased to be solely a matter of individual choice and became a domain subject to strong social influence. It became a subject

of media discourse and a component of market strategy or even manipulation. Researchers often use the term "subjective well-being". The term "well-being" appeared in the definition of health proposed by the WHO in 1948 as "Health is a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity" (*Constitution of the World Health Organization* 2025).

As early as 1890, William James, in his classic work *The Principles of Psychology* (James 1890), wrote about the "stream of consciousness" and the spiritual sources of well-being – these are some of the first psychological approaches to the subject of happiness. However, the 1990s, Martin Seligman (2011) proposed a new direction of research focused on mental well-being, which is considered to be the true beginning of positive psychology. The first theoretical models describing the structure of happiness were developed at that time. The first of these is the PERMA model (Positive emotions, Engagement, Relationships, Meaning, Accomplishment), which identifies five key pillars of well-being. Another theory is Mihály Csíkszentmihályi's flow theory (Csikszentmihalyi 1990), which characterizes happiness as a state of complete engagement and absorption in a given activity. It also gained popularity in the 1990s but had been developed since the 1970s.

Along with the development of positive psychology, empirical research on happiness, life satisfaction, and well-being has also progressed (Czapiński 1994). More and more factors that influence well-being, both individual and social, have begun to be taken into account. This issues has increasingly attracted interdisciplinary attention and "become a significant topic for empirical studies" (Hallberg and Kullenberg 2019, p.42) situating questions of personal destiny and worldview within a global framework.

Since 2012, the annual World Happiness Report (World Happiness Report, n.d.), Gallup (2024), the UN Sustainable Development Solutions Network, and an independent editorial team have been involved in the creation of the reports.

The second most important report is the Better Life Index, developed by the OECD in 2011 (*OECD Better Life Index,* n.d.). It focuses on four key areas: environmental sustainability, improving well-being, reducing inequality, and system resilience. The OECD initiative takes the form of an interactive portal where users can assign weights to individual categories themselves to see which country best meets their expectations. What can we learn from the report? First of all, that the study covers areas of socio-economic progress (follow in OECD): "1. Housing: housing conditions and costs (e.g. real estate pricing). 2. Income: household income (after taxes and transfers) and net financial wealth. 3. Jobs: earnings, job security and unemployment. 4. Community: quality of social support network. 5. Education: education and what one gets out of it. 6. Environment: quality of environment (e.g. environmental health). 7. Governance: involvement in democracy. 8. Health. 9. Life Satisfaction: level of happiness. 10. Safety: murder and assault rates. 11. Work–life balance" (*OECD Better Life Index,* n.d.).

The input data are translated into approximately 80 indicators of high socio-economic utility, useful in management as well as in the creation and modeling of political activities. After reviewing the above thematic areas, it can be concluded that happiness only concerns the ninth item. What constitutes the overall picture of the study is better described as satisfaction. It is difficult to find direct references to ancient concepts of happiness in the proposed areas. The hedonistic perspective seems to be the closest, while the Aristotelian concept of eudaimonia seems to be the most distant. The indicators focus on quick responses,

measurable states, and satisfaction, which makes it difficult to appreciate long-term character formation, where progress can be difficult to describe.

The third example is the Happy Planet Index (*Happy Planet Index,* n.d.), an indicator developed by the British think tank New Economics Foundation, which has been measuring well-being in individual countries since 2006, paying particular attention to environmental protection and the fact that the goal of economic development should be to ensure people's health and happiness. The HPI calculations focus on factors such as ecological footprint and sustainable development. The index is one of the most compelling calculations supporting the foundation's goals, which can be summarized as follows: "1. A Fast and Fair Transition. Build a fast and fair transition to a green economy through a deep and equitable transformation to address the climate crisis. 2. Guarantee Economic Security for all. Guarantee economic security for all by rebuilding effective and sustainable public services to give everyone access to life's essentials. 3. Shift Power to People and Communities. Shift power to people and communities through democratic ownership of the economy and devolving political power" (*Happy Planet Index,* n.d.).

The last of the selected examples is Gallup Global Emotions – an annual report prepared by an American international analytical and consulting company (*Gallup Global Emotions Report* 2024). It serves as a measure of emotions, providing information on the state of happiness understood as the sum of positive emotions minus the sum of negative emotions, examining them through surveys conducted worldwide in 100 countries.

Below is a table listing the most popular reports on the subject of happiness.

Table 2. Selected happiness and well-being indices

Name	Organization	Description
World Happiness Report	UN	The most important ranking based on subjective well-being
Better Life Index	OECD	Interactive, multiple dimensions of quality of life
Happy Planet Index	NEF (UK)	A combination of happiness and ecology
Gallup Global Emotions	Gallup, Inc.	Everyday emotions — happiness "here and now"

In this way, we are embarking on a new path to happiness on a global scale. To summarize the types of indicators, they can be compared with the ideas of ancient concepts of happiness.

The WHR and Better Life Index present happiness as living in decent conditions, freedom from fear, and a sense of security, attempting to describe a number of external components of happiness in line with the visions of Aristotle and Epicurus. The HPI, which values harmony, especially living in harmony with the environment, refers to the philosophies of the Stoics and Epicureans. GGE, on the other hand, corresponds most closely to the philosophy of the Cyrenaics, leaning towards the emotional "here and now".

It is also worth mentioning how apps dedicated to happiness or well-being work.

One of the most popular apps related to happiness and well-being is Headspace[4], which has over 70 million downloads and reaches over 100 million users worldwide. One of its founders is a former Buddhist monk who wanted to popularize the practice of meditation. The company is currently valued at approximately $3 billion, with revenues of $348.4 million in 2024 (Latka 2024). The app supports meditation and mindfulness practices rooted in

4. headspace.com

Eastern philosophies, but it also contains elements close to Stoicism – especially distancing oneself from emotions, accepting fate and things over which we have no control. There are no direct references to ancient philosophy in it. It is focused on personalization and analysis of individual predispositions (frequency of use, specific time, matching breathing exercises to biometric results).

Another app is Calm[5], which had approximately 4.5 million paying subscribers in 2023 and an estimated 180 million downloads (Curry 2025). It is one of the longest-running apps in the category and offers meditation practices as well as tools for relaxation, visualization, creativity, and concentration. The app can be personalized to user preferences, adapts content to a user's current emotional state, and can even flag a potential risk of depression.

These two apps reportedly account for about 90% of the meditation-app market and are among the most popular results for apps associated with the keyword "happiness". Due to the significant difference in popularity and reach of the apps, it can be assumed that the next ones will be sufficiently popular to be selected among those that occupy a smaller share of the app market.

The third app is Insight Timer[6], also dedicated to free meditation. It gives you access to thousands of free guided meditations. It offers resources for stress reduction and develops mindfulness, support for sleep problems, and a wide selection of meditation music. The app encourages regular practice, allows us to participate in live classes, and emphasizes community building. It has an extensive content database, the selection of which is supported by algorithms. Work is also underway on solutions in which a "virtual mentor" guides users through the practice.

The fourth app discussed is Happify[7]. It offers a set of exercises and games aimed at strengthening mental health and improving well-being. It draws most heavily on the principles of positive psychology but also refers to behavioral research and mindfulness practices. This approach can be seen as an attempt to follow the philosophy of Aristippus of Cyrene, which emphasizes the role of pleasure and mood improvement as a path to a happy life. The app also uses artificial intelligence for clinical analysis and evaluation of the effectiveness of the proposed therapeutic paths.

This concise overview of the analyzed applications shows that we are dealing with behaviors preferred by customers. The focus on quick results does not always encourage actions with delayed gratification, and happiness as the result of a long-term process is a less attractive proposition and a "commodity" that is more difficult to sell.

At different stages of life, people subordinate themselves to the pursuit of goals, believing that the achievement of them will give them a better state than the one they are currently in (Uribe et al. 2023). They do this because of their accepted vision of their lives. This does not always mean full compliance with the imagined vision of happiness but usually reflects a certain stage or state that provides the circumstances for achieving a state of greater happiness.

It can be noted that it is extremely difficult to define a set of actions that ensure happiness. There is confusion regarding minor intermediate goals and interpersonal obligations, an imbalance between the realization of one's own goals and the fulfillment of expectations

5. calm.com
6. insighttimer.com
7. happify.com

resulting from the socio-cultural model in which we function. Achieving happiness is therefore the result of many factors, and the search of individuals for a properly formulated concept of happiness is nothing more than their waiting for instructions and a kind of recipe. Technological progress brings opportunities and threats, including formulas for happiness that are directed at the individual client and, the adaptation of mental health support (Linardon et al. 2024; Uribe et al. 2023). With only a rare need to visit a specialist, the expansion of offers with new content, health monitoring, and the possibility of saving costs seems inviting.

Not only do we believe in new technological achievements, but we are beginning to perceive artificial intelligence as a kind of authority (Hauswald 2025), highly valuing the ability to describe reality mathematically, sometimes without reference to the depth of human experience and unique individuality.

Summary

The 21st century has seen an intensification of research into happiness, particularly its quantitative measurement. On the modern path to happiness, people may encounter linguistic transformations created by AI operating on a processed description of reality. Ancient concepts remain an inspiration, but under the influence of modern tools, they become simplified and optimized. In the proposed applications, we focus on avoiding undesirable experiences and maximizing short-term sensations – because they determine whether we will return to the application. Happiness is reduced to measurable, mathematical aspects of the phenomenon. The depth of human experience, the mystery of existence and individuality, which is impossible to present in applications, seems sometimes even more inaccessible. On the other hand, it should be noted that these applications enable a wide audience to work on them and perhaps contribute their development, for they are accessible to many people anywhere and anytime, sometimes enabling persons to take first step towards improving their well-being and supporting therapeutic activities.

The temporal dimension also remains an important theme – contemporary approaches balance between quick results and the long-term process of striving for happiness. It is to be hoped that the measurements will contribute to finding a balance between the hedonistic approach and eudaimonia (Schmitz, Burk, and Wiese 2025).

Questions arise about the role of technological advances and about the growing perception of AI as a primary source of knowledge, even including knowledge about one's own happiness – as well as about the risks of technological dependence and the problem of shared responsibility for actions taken under its influence.

The article raises many issues that require broader, more in-depth study, appropriate to the scope and importance of the content addressed.

References

Angner, Erik. 2011. The Evolution of Eupathics: The Historical Roots of Subjective Measures of Wellbeing. *International Journal of Wellbeing* 1 (1): 4–41. https://doi.org/10.5502/ijw.v1i1.14.

Arystoteles. 1996. *Dzieła wszystkie. Tom 5.* Translated by Daniela Gromska. Warszawa: PWN.

Constitution of the World Health Organization. 2025. World Health Organization. https://www.who.int/about/governance/constitution accessed August 15, 2025.

Csikszentmihalyi, Mihaly. 1990. *Flow: the psychology of optimal experience.* New York: Harper & Row.

Curry, David. 2025. *Calm Revenue and Usage Statistics.* https://www.businessofapps.com/data/calm-statistics accessed August 15, 2025.

Czapiński, Janusz. 1994. *Psychologia szczęścia. Przegląd badań i zarys teorii cebulowej.* Warszawa: Pracownia Testów Psychologicznych PTP.

Fredrickson, Barbara. 2002. Positive emotions. In *Handbook of positive psychology,* edited by C.R. Snyder and Shane Lopez, 120–134. Oxford: Oxford University Press.

Gallup Global Emotions Report. 2024. Gallup. https://www.gallup.com/analytics/349280/gallup-global-emotions-report.aspx accessed July 15, 2025.

Hallberg, Margareta, and Christopher Kullenberg. 2019. Happiness Studies: Co-Production of Social Science and Social Order. *Nordic Journal of Science and Technology Studies* 7 (1): 42–50. https://doi.org/10.5324/njsts.v7i1.2530.

Happy Planet Index. n.d. New Economics Foundation. https://happyplanetindex.org/ accessed August 15, 2025.

Hauswald, Rico. 2025. Artificial Epistemic Authorities. *Social Epistemology* 39 (6): 716–725. https://doi.org/10.1080/02691728.2025.2449602.

Haybron, Dan. 2007. Well-Being and Virtue. *Journal of Ethics and Social Philosophy* 2 (2): 1–28. https://doi.org/10.26556/jesp.v2i2.21.

James, William. 1890. *The Principles of Psychology. Vol. 1.* New York: Henry Holt / Company.

Karaś, Dominika. 2019. Pojęcia i koncepcje dobrostanu: przegląd i próba uporządkowania. *Studia Psychologica* 19 (2): 5–23. https://doi.org/10.21697/sp.2019.19.2.01.

Krasucka, Agata. 2014. Szczęście człowieka jako problem filozoficzno-psychologiczny na przestrzeni dziejów. *Studia Paedagogica Ignatiana* 17:133–146. https://doi.org/10.12775/SPI.2014.007.

Latka, Nathan. 2024. *How Headspace hit $348.4M revenue and 100M customers in 2024.* https://getlatka.com/companies/headspace accessed August 15, 2025.

Linardon, Jake, Mariel Messer, Simon B. Goldberg, and Matthew Fuller-Tyszkiewicz. 2024. The efficacy of mindfulness apps on symptoms of depression and anxiety: An updated meta-analysis of randomized controlled trials. *Clinical Psychology Review* 107:102370. https://doi.org/10.1016/j.cpr.2023.102370.

Maslow, Abraham. 1954. *Motivation and Personality.* New York: Harper & Brothers.

Mitchell, Lewis, Morgan R. Frank, Kameron Decker Harris, Peter Sheridan Dodds, and Christopher M. Danforth. 2013. The Geography of Happiness: Connecting Twitter Sentiment and Expression, Demographics, and Objective Characteristics of Place. *PLoS ONE* 8 (5): e64417. https://doi.org/10.1371/journal.pone.0064417.

Mordarski, Ryszard. 2023. Koncepcja felicytologii Władysława Tatarkiewicza na tle współczesnych filozofii szczęścia. *Przegląd Filozoficzny. Nowa Seria* 32 (1): 255–274. https://doi.org/10.24425/pfns.2023.146792.

OECD Better Life Index. n.d. OECD. https://www.oecd.org/en/data/tools/oecd-better-life-index.html accessed July 15, 2025.

Richter, Caroline G., Celine Mylx Li, Adam Turnbull, Stephanie L. Haft, Deborah Schneider, Jie Luo, Denise Pinheiro Lima, Feng Vankee Lin, Richard J. Davidson, and Fumiko Hoeft. 2024. Brain imaging studies of emotional well-being: a scoping review. *Frontiers in Psychology* 14:1328523. https://doi.org/10.3389/fpsyg.2023.1328523.

Schmitz, Bernhard, Christian L. Burk, and Bettina S. Wiese. 2025. Enhancing Life Satisfaction through Eudaimonic, Hedonic, and Combined Interventions: New Training Approaches Relevant to Theory and Practice. *Journal of Happiness Studies* 26 (4): 60. https://doi.org/10.1007/s10902-025-00872-w.

Seligman, Martin E.P. 2011. *Flourish: A visionary new understanding of happiness and well-being.* New York: Free Press.

Seneca. 1917. *Epistles, Volume I: Epistles 1-65.* Translated by Richard M. Gummere. Cambridge: Harvard University Press.

Swoboda, Antoni. 1997. Epikurejskie rozumienie przyjemności i szczęścia. *Roczniki Filozoficzne* 45-46 (2): 69–82.

Tatarkiewicz, Władysław. 2005. *O szczęściu.* Warszawa: PWN.

Uribe, Fabio Alexis Rincón, Maria Fernanda Monteiro Favacho, Paula Marília Nascimento Moura, Diana Milena Cortés Patiño, and Janari Da Silva Pedroso. 2023. Effectiveness of an app-based intervention to improve well-being through cultivating positive thinking and positive emotions in an adult sample: study protocol for a randomized controlled trial. *Frontiers in Psychology* 14:1200960. https://doi.org/10.3389/fpsyg.2023.1200960.

White, Nicholas. 2008. *Filozofia szczęścia: od Platona do Skinnera.* Translated by Marek Chojnacki. Kraków: WAM.

World Happiness Report. n.d. *World Happiness Report.* https://www.worldhappiness.report/ accessed July 15, 2025.

Philosophical Problems in Science (2025), **79**, 59–71
(Zagadnienia Filozoficzne w Nauce)
doi:10.59203/zfn.79.727

ARTICLE

From simple rules to quasi intentionality: emergent coordination and collective cognition

Anna Sarosiek

University of the National Education Commission, Poland
Email: anna.sarosiek@uken.krakow.pl

Abstract

This article examines whether quasi-intentional organisation characterised by anticipation, adaptation and retention can arise in distributed systems composed of simple agents without global goal representation. The analysis draws on observations of *Camponotus barbaricus* ants that construct and stabilise pine-needle barriers in the absence of central control. An agent-based model implemented in NetLogo formalises selected construction dynamics through local rules governing exploration, material deposition and reinforcement of stable configurations. Three operational indicators are introduced to examine system-level organisation: pre-disturbance increases in construction activity, directional changes in structural stability across iterations and reconstruction of previously stabilised patterns after partial removal. The model functions as a constrained artificial-life environment for analysing purposive-looking organisation emerging from distributed interaction and environmental traces. The concept of quasi intentionality is situated within biosemiotic accounts of environmental semiosis and a relational ontology of agency, and is discussed in relation to current approaches to symbiotic and beneficial artificial systems.

Keywords: quasi intentionality, distributed systems, biosemiotics, stigmergy, artificial life, agent-based modelling, symbiotic AI

1. Introduction

The notion of intentionality underpins classical conceptions of mind and agency. The term intentionality was introduced by Franz Brentano (2014), who defined it as the mind's capacity for being about something, that is, the ability to represent or refer to objects, states and events. In philosophical literature, intentionality is often identified with the possession of mental content and conscious intentions (Searle 1983; Dennett 1987). Although many alternative accounts exist, from pragmatic through phenomenological to enactive, this article adopts the classical Brentano–Dennett definition as a point of reference. The aim is not to reconstruct the entire terminological debate but rather to analyse quasi intentional behaviours in distributed systems where the absence of central control and explicit goal representation poses a fundamental challenge to classical conceptions of intentionality.

Algorithms inspired by nature, such as Particle Swarm Optimisation (PSO) and Ant Colony Optimisation (ACO), demonstrate that complex optimisation behaviours can emerge from simple local rules. In PSO, agents move through the solution space by tracking the best global and local positions without any need for central coordination (Kennedy and Eberhart 1995). In ACO, indirect coordination through environmental traces known as stigmergy enables the discovery of shortest paths and the construction of complex structures (Bonabeau, Dorigo, and Theraulaz 1999; Dorigo and Stützle 2004). Studies of self-organisation in biological systems confirm that global patterns can arise entirely from iterations of simple behaviours responding to local information (Camazine et al. 2003). These findings suggest that coordinated and apparently goal-directed patterns may emerge from distributed interactions even in the absence of central representation or explicit planning.

These findings challenge the anthropocentric assumption that purposive behaviour requires conscious intention. Ant colonies can build and refine structures even though individual workers lack any global representation of the whole. Classic research on stigmergy has shown that social insects leave traces in the environment (pheromones or arranged pine needles) that guide the actions of other individuals and enable collective coordination without direct communication (Deneubourg and Goss 1989; Bonabeau, Dorigo, and Theraulaz 1999; Camazine et al. 2003).

One of the key challenges in contemporary artificial intelligence is the design of beneficial systems that adaptively cooperate with humans and the environment. Traditional approaches rely on explicit programming of goals and utility functions. The Symbiotic Beneficial AI framework (Polak and Krzanowski 2023) proposes an alternative: systems in which beneficial patterns emerge through local interactions modelled on biological symbiosis. This approach shifts emphasis from centralised optimisation to the co-evolution of cooperation and adaptation within dynamic cognitive ecosystems, where artificial agents function as participants rather than merely as tools.

Empirical inspiration for this work comes from observations of *Camponotus barbaricus* constructing pine needle barriers (Franks et al. 1992; Camazine et al. 2003). In laboratory settings, workers initiate construction in the absence of immediate threats, refine the structure over successive iterations (Deneubourg and Goss 1989) and consolidate effective patterns after barrier removal (Camazine et al. 2003). This article develops a conceptual and modelling framework for analysing quasi-intentional organisation in distributed systems. An agent-based simulation, drawing on selected construction dynamics of *Camponotus barbaricus*, functions here as a formal and illustrative model used to examine whether patterns corresponding to anticipation, adaptation and retention can arise from local interaction rules in the absence of central representation.

2. Quasi intentionality as an intermediate category

Classical theories of intentionality (Brentano 2014; Dennett 1987; Searle 1983) assume an inseparable link between intentionality, mental content and consciousness. Intentional being about something requires representation and conscious purpose, yet in distributed systems lacking a central representational unit, functionally analogous behaviours to purposeful action are observed. To describe these phenomena, the notion of quasi intentionality is introduced, understood here as system-level purposive organisation that emerges at the level of the entire

system in the absence of central representation or conscious intention. Quasi intentionality is treated here as a lower-level form within a graded spectrum of intentional organisation, occupying a position between basic regulatory purposiveness and fully representational, conscious intentionality.[1] Three basic system functions are distinguished:

1. Anticipation
   The system initiates adaptive actions before a clear threat stimulus appears. Anticipation consists in the reduction of surprise relative to recurrent perturbations. In representational systems this may involve internal models of future states (Friston 2010; Clark 2016); in quasi intentional systems, analogous effects may arise from emergent network dynamics and distributed environmental feedback (Camazine et al. 2003).
2. Adaptation
   The system modifies action strategies over successive iterations on the basis of global and local feedback. In complex systems theory, adaptation occurs when local rules change in response to the state of the whole system, as described by positive and negative feedback mechanisms (Haken 1977; Prigogine and Stengers 1984).
3. Retention
   The system retains and reinforces effective patterns via distributed memory. In Hopfield networks, the system state functions as an information carrier (Hopfield 1982), and in insect colonies, durable environmental structures store cues that guide subsequent actions (Deneubourg and Goss 1989).

The choice of these three indicators rests on their fundamental role in a functional account of purposiveness. In this framework, anticipation, adaptation and retention are treated as operational descriptors of system-level organisation rather than as claims about internal mental states or representational content. This triad of forecast, correction and consolidation constitutes the core of any form of purposeful action, whether of individual agents or as an emergent system trait. Integration of these components enables description of quasi intentionality in collective systems even when individual units lack internal representations or conscious goals.

Quasi intentionality avoids anthropomorphism by allowing discussion of system-level intention without attributing consciousness to individual elements. This aligns with the enactive conception of cognition (Varela 1992; Thompson 2007), according to which intelligence arises from dynamic interaction between organism and environment rather than from internal representations. It is also consistent with active externalism (Clark and Chalmers 1998), which treats environment and artefacts as participants in cognitive and agency processes.

Recent work on minimal cognition further supports the view that purposive organisation can arise without central representation or conscious planning. Research on basal cognition and distributed regulation argues that even simple biological systems exhibit forms of goal-directed organisation grounded in metabolic and sensorimotor coupling rather than in

1. The term *quasi intentionality* is introduced here as a working analytical category situated within a continuous line of research on minimal and distributed forms of purposive organisation. The approach draws on Uexküll's conception of organism–environment relations and functional meaning within the *Umwelt* (Uexküll 1921, 1934), on biosemiotic accounts of regulation and sign-mediated interaction (Hoffmeyer 2011; Kull et al. 2009), and on enactive theories of minimal agency and sense-making (Varela 1992; Thompson 2007; Barandiaran, Di Paolo, and Rohde 2009). Within this framework, these perspectives are treated not as competing approaches but as complementary accounts of purposive organisation emerging without central representation.

symbolic representation (Lyon 2015; Levin 2019). From this perspective, cognition is not restricted to brains but extends across regulatory processes that maintain system integrity and guide adaptive behaviour. Such approaches shift attention from internal representations to patterns of coordination across bodies, environments and interaction histories. The present notion of quasi intentionality remains more restrained: it does not attribute cognition in a strong biological sense to artificial or collective systems, but it does draw on this line of work to support the claim that purposive organisation can be analysed at the level of system dynamics without presupposing explicit representation or awareness.

A complementary line of argument is developed in dynamical and interactivist accounts of minimal agency. Dynamical systems approaches emphasise how coherent behavioural patterns can emerge from the continuous coupling of agents and environments without reliance on symbolic control structures (Beer 2014). Interactivist theories similarly argue that normative organisation and goal-directed behaviour can arise from processes that maintain and regulate their own conditions of viability, even in the absence of internally represented goals (Bickhard 2009). These perspectives reinforce the interpretation adopted here: quasi intentionality designates a level of organisation at which coordinated, purposive patterns can be identified in the dynamics of a system as a whole, without implying that individual components possess intentions, representations or conscious awareness.

3. Relational ontology, semiosis and symbiotic AI

The concept of quasi intentionality presupposes a shift from substance ontology towards a relational ontology in which agency emerges from networks of interaction rather than residing in individual entities. According to autopoietic theory (Maturana and Varela 1980), living systems maintain their identity through continuous exchanges with their environment. In biosemiotic approaches this process is understood as semiosis, in which each interaction creates and interprets meaningful traces in the environment, thereby constituting the subjective *Umwelt* of organisms (Uexküll 1921; Kull et al. 2009; Hoffmeyer 2011). In quasi intentional systems, semiosis plays a key role in retention, as durable environmental signs record historically effective patterns, and in adaptation, as those signs guide subsequent adjustments of behaviour.

In the context of artificial systems, the use of biosemiotic terminology should be understood cautiously. Following discussions in artificial life and cognitive robotics, semiotic and intentional vocabularies are often employed as analytic and modelling tools rather than as ontological claims about artificial agents themselves (Emmeche 1994; Ziemke 2001). Within this perspective, semiosis does not imply the presence of intrinsic meaning or interpretation in a human or biological sense. Instead, it designates a relational configuration in which environmental modifications constrain subsequent system behaviour and thereby acquire functional significance within ongoing dynamics. The present framework adopts this restrained usage: biosemiotic concepts are used to characterise patterns of interaction and stabilisation in distributed systems, not to attribute full-fledged semiotic capacities to individual agents.

Within this relational ontology, higher levels of system organisation may influence the behaviour of components. This mechanism, often described as downward causation (Campbell 1974), can be understood in terms of the formation of dynamic attractors. Such attractors organise local agent behaviour so that the system as a whole achieves functional

outcomes despite the absence of a central plan (Barandiaran, Di Paolo, and Rohde 2009). Research on minimal forms of agency indicates that even simple self-sustaining and boundary-regulating systems can exhibit emergent forms of agency (Barandiaran, Di Paolo, and Rohde 2009; Thompson 2007).

Quasi intentional mechanisms operate via self-organisation and stigmergy. Self-organisation is a process in which ordered patterns arise from repeated local interactions governed by simple rules without external control (Haken 1977; Prigogine and Stengers 1984). In a biosemiotic framing, each act of construction, whether the placement of a needle or the deposition of pheromone, may be treated as an instance of semiosis that serves as a stimulus for subsequent actions. Consequently, global patterns emerge stigmergically: environmental traces guide agent behaviour while functioning as meaning carriers that encode and stabilise historically effective patterns within the system. Such environmental memory and meaning carriers require a balance of positive feedback that reinforces success and negative feedback that prevents uncontrolled growth. This balance enables the system both to explore new solutions and to stabilise effective patterns.

From a modelling perspective, the notion of semiosis can be partially operationalised by specifying how environmental traces are generated, maintained and differentially weighted in subsequent interactions. Work in artificial life and robotics has emphasised that semiotic descriptions of artificial systems are most defensible when linked to explicit mechanisms of coupling between agent activity and environmental modification (Emmeche 1994; Ziemke 2001). In this sense, semiosis is not treated as an additional property layered onto the model, but as a way of describing how distributed memory and constraint structures function within system dynamics. The present model therefore treats environmental traces as operational carriers of constraint and coordination, allowing biosemiotic terminology to be used in a controlled and explicitly defined manner.

The Symbiotic Beneficial AI framework (Polak and Krzanowski 2023) draws on these inspirations. In nature, mutualistic relationships, such as mycorrhizal associations or cleaning symbioses, demonstrate that complex beneficial functions can emerge through local semiosis and mutual tuning of partners (Thompson 2007; Hoeksema and Bruna 2015). In artificial systems, semiosis may be translated into mechanisms that enable agents to create and interpret environmental traces in digital environments, thereby giving rise to emergent cooperation and adaptation. Unlike economic or consensus-based algorithms, symbiotic AI does not assume optimisation of a single utility function but instead seeks to create conditions for the emergence of complementary behaviours within dynamic cognitive ecosystems, where agents function as participants in the cognitive process.

4. Biological inspiration and translation to an ALife model

Observations of the ant *Camponotus barbaricus* provide empirical inspiration for analysing quasi intentional mechanisms in collective systems. In controlled formicarium settings, workers are observed to build pine-needle barriers even in the absence of immediate threats, corroborating earlier experimental and field reports on stigmergic construction behaviour in social insects (Franks et al. 1992; Deneubourg and Goss 1989; Camazine et al. 2003). Initial construction phases can be described in functional terms as anticipatory, insofar as workers gather material and begin forming a protective structure prior to overt perturbations. Subsequent iterations

of the building process exhibit adaptive adjustments in structural stability and assembly speed through local reinforcement of successful configurations. Rather than depositing needles randomly, workers tend to concentrate effort on structurally critical sections of the barrier (Deneubourg and Goss 1989; Bonabeau, Dorigo, and Theraulaz 1999). Following partial removal of the structure, colonies frequently reconstruct previously stabilised segments first, even when different workers participate in the rebuilding (Camazine et al. 2003). These patterns of anticipation, adaptation and retention provide a biological reference point for analysing purposive organisation in systems lacking central representation.

At the level of individual behaviour, workers do not possess a global representation of the structure being built and do not operate with an explicit plan of the final configuration. Each agent responds only to local material densities, spatial constraints and traces left by previous actions. Nevertheless, colony-level dynamics can give rise to stable and functionally organised structures that are initiated prior to disturbances, modified across successive iterations and partially reconstructed after removal. This discrepancy between strictly local rule-following and coherent global organisation has long been recognised as a defining feature of stigmergic systems and collective self-organisation (Deneubourg and Goss 1989; Bonabeau, Dorigo, and Theraulaz 1999; Camazine et al. 2003).

The theoretical interest of such systems lies precisely in this tension between local responsiveness and global coherence. Without invoking internal plans or representations at the level of individual agents, the colony exhibits coordinated dynamics that can be described functionally as anticipatory, adaptive and retentive. Similar forms of system-level organisation have been discussed in research on minimal and distributed agency, where purposive structure is treated as emerging from ongoing interaction and constraint rather than from centrally represented goals (Barandiaran, Di Paolo, and Rohde 2009; Thompson 2007). Within this perspective, quasi intentionality can be used as a descriptor of system-level organisation that arises from distributed interactions and environmental feedback, without attributing full-fledged intentional states to individual components.

In an agent-based simulation implemented in NetLogo (Wilensky 1999), each agent executes three basic local rules corresponding to the behavioural patterns described above: exploration and collection of resource needles, deposition of material near existing accumulations, and reinforcement of stable areas while avoiding unstable ones. A functional analogue of anticipation is modelled through an early-activation threshold that causes agents to respond to local material density rather than to an explicit threat signal. Adaptation is implemented by dynamically increasing the attractiveness of sites where previously deposited material has remained stable across iterations. Retention emerges through the persistence of deposition–attractiveness fields that function as a form of distributed environmental memory within the model.

5. Methodology and Model Architecture

The model is presented as a constrained artificial-life system designed to explore whether quasi-intentional patterns may arise from explicitly specified local interaction rules. Rather than reproducing biological colonies in full detail, it formalises selected construction dynamics observed in *Camponotus barbaricus* and evaluates whether system-level patterns corresponding to anticipation, adaptation and retention can emerge from these local interactions. It is

intended as a formal and illustrative framework for examining distributed organisation under controlled conditions, not as an empirical reproduction of biological colonies.[2]

An artificial life simulation was implemented in NetLogo (Wilensky 1999) on a two-dimensional grid measuring 30×30 cells. The grid represents a simplified forest environment with a central nest at coordinates (15, 15) and resources (food and pine needles) distributed randomly. The initial population comprises ten worker agents located within the nest. Agent reproduction occurs when nest reserves exceed fifty units of food, consuming thirty units per new agent and thereby introducing simple demographic dynamics. Agents operate autonomously, each maintaining its own coordinates, behavioural state (foraging / transport / construction) and the capacity to carry a single resource (food or pine needle). Perception is local and limited to a radius of three cells, introducing a constraint on information access comparable to those reported in biological observations (Camazine et al. 2003; Franks et al. 1992). No agent has access to global structural information or to any representation of forthcoming environmental disturbances.

Agent behaviour is governed by three sets of local rules derived from and abstracting selected construction dynamics described in biological observations and literature (Deneubourg and Goss 1989; Bonabeau, Dorigo, and Theraulaz 1999; Camazine et al. 2003):

1. Foraging
   Agents perform a correlated random walk, turning up to $\pm 45^\circ$ with probability 0.7 or moving straight otherwise. Upon detecting food within perception range (radius of three cells), agents pick it up with probability 0.8 and return to the nest following a pheromone gradient that decays at rate 0.95 per tick.
2. Pine needle collection
   During foraging, agents encountering pine needles pick them up with probability

$$P_{\text{collect}}(d) = 0.3e^{-0.1d},$$

   where d denotes Euclidean distance to the nearest nest entrance. This formulation yields the highest collection probability near the entrance and decreases it with distance, reflecting the tendency to concentrate construction near critical entry regions, as observed in stigmergic construction processes (Deneubourg and Goss 1989).
3. Barrier construction
   Agents carrying a needle deposit it with probability

$$P_{\text{deposit}}(\rho) = 0.1 + 0.4 \left(\frac{\rho}{\rho_{\max}} \right)^2,$$

   where ρ represents local needle density computed as the number of needles within a Moore neighbourhood of radius one (eight surrounding cells) around the agent's current cell, and $\rho_{\max}$ denotes the maximum observed value of ρ within the current simulation run, used for normalisation. If $\rho_{\max} = 0$, deposition defaults to the baseline probability $P_{\text{deposit}} = 0.1$.

2. The full simulation model (NetLogo 7.0.3) is available for review via an anonymous Open Science Framework repository: https://osf.io/srwfh/?view_only=264c47c847b1460a9c2e382448756589.

After deposition, the stability of that position is monitored. Stability is defined as the persistence of at least one needle at that location for consecutive ticks. If stability persists for ten consecutive ticks, the attractiveness of that site is increased by a factor of 1.2; for each additional completed block of ten stable ticks (i.e. at 20, 30, 40 ticks, etc.), the same reinforcement is applied again. If the number of needles at that location falls to zero, the site is treated as unstable and attractiveness is reduced. Stability is implemented as a patch-level persistence variable rather than an agent-level memory. Reinforcement therefore operates through environmental modification rather than internal state retention.

Environmental resources regenerate dynamically, with pine needles appearing at probability 0.02 per cell per tick (excluding the nest) and food at probability 0.01 per empty cell per tick. The environment includes designated zones: a storage area (5×5 cells at the centre), an expanded entrance zone (four cells on the nest perimeter), and a nursery area for reproduction. Modifications to the environment (namely needle deposition) bias agent movement by altering local attractiveness fields, thereby influencing subsequent deposition patterns.

To examine structural robustness under disturbance, controlled perturbations are introduced at regular intervals. When enabled, a perturbation occurs every 500 ticks and consists of a uniform 50% reduction of structural values across the environment. Each interval between successive perturbations defines a construction cycle. This cyclical disturbance allows analysis of cross-cycle changes in structural stability and the degree of spatial overlap between pre- and post-perturbation configurations.

Data generated within the simulation were inspected to characterise general activity patterns and system responses to parameter changes. Analytical procedures were used in an exploratory and illustrative manner to identify shifts in building dynamics and to assess the internal consistency of the operational measures introduced above. The purpose of these analyses is not confirmatory testing, but the clarification of how the proposed indicators of anticipation, adaptation and retention function within the model environment.

6. Operationalisation of quasi-intentionality

Quasi-intentionality is operationalised here as a minimal regime of norm-guided, historically sensitive organisation detectable at the level of system dynamics. The indicators below do not attribute representational content to agents. Rather, they measure whether collective dynamics exhibit temporal asymmetry, cross-cycle stabilisation, and structured reconstruction under controlled perturbation.

All values reported below are based on $N = 300$ independent runs with varying random seeds.

1. Temporal asymmetry (anticipatory activation)
   Building activity was quantified as the number of structural reinforcement events per tick, aggregated over 100-tick windows preceding perturbation. Across runs, mean pre-perturbation activation was

$$M = 0.492, \quad SD = 0.12.$$

This indicates a stable forward-directed bias in structural mobilisation prior to disturbance. The magnitude of the effect remains moderate and does not reach the previously defined conservative threshold of $z > 2$. It is therefore interpreted as persistent temporal asymmetry rather than strong predictive escalation.

2. Adaptation (cross-cycle stabilisation)
   Structural stability was measured at the patch level via a persistence counter (stable-ticks). For each cycle, a cycle-level stability score was computed as the mean stability of structurally active patches.
   Mean early-cycle stability was

$$M_{\text{early}} = 790, \quad SD \approx 190,$$

   whereas mean late-cycle stability increased to

$$M_{\text{late}} = 7086, \quad SD \approx 950.$$

   This corresponds to an approximately 8.97-fold increase across cycles. The correlation between cycle index and stability remained below $r = 0.7$, indicating that the increase is directional but non-linear. The effect reflects progressive structural consolidation under repeated disturbance rather than simple linear accumulation.
3. Retention (historical persistence)
   Retention was computed as spatial overlap between structural configurations before and after perturbation. Across runs, mean retention was

$$M = 0.80, \quad SD = 0.07,$$

   with 73% of runs exceeding the conservative threshold of $R > 0.7$.
   This indicates strong, though not deterministic, reconstruction of previously stabilised configurations.

Taken together, these indicators demonstrate a regime in which collective dynamics exhibit measurable temporal asymmetry, cumulative stabilisation across cycles, and historically mediated reconstruction. The reported values describe tendencies within the present parameter configuration and do not claim universality across all possible parameter settings.

7. Discussion

The simulation generates three measurable patterns from explicitly specified local probabilistic rules: temporal asymmetry in construction activity (M = 0.492, SD = 0.12), cross-cycle structural consolidation (8.97-fold stability increase from early to late cycles), and spatial reconstruction following perturbation (M = 0.80, SD = 0.07). These patterns arise without central coordination and without global goal representation.

The magnitude of these effects constrains their interpretation. Pre-disturbance activation remains below the conservative threshold ($z > 2$), indicating a weak but persistent forward bias rather than strong predictive escalation. Cross-cycle stability correlates with cycle index at $r < 0.7$, suggesting directional but non-linear consolidation. Retention exceeds $R > 0.7$ in 73% of runs, demonstrating robust yet non-deterministic reconstruction. The reported

values therefore describe stable tendencies within the present parameter configuration rather than universal behavioural laws.

Importantly, it is the co-occurrence of forward temporal bias, cumulative stabilisation and selective reconstruction that defines the observed regime. The results do not demonstrate predictive modelling or representation. They demonstrate a constrained dynamic organisation in which past stabilisation and present reinforcement jointly shape the distribution of future structural states.

Three limitations bound the scope of generalisation. First, the environment is simplified and two-dimensional; agents are homogeneous; resource regeneration is fixed. Biological colonies operate under heterogeneous terrain, agent differentiation and stochastic perturbation. Second, although N = 300 runs establish internal consistency, no direct parameter calibration against empirical ant data has been performed. Third, the indicators introduced (anticipation, adaptation, retention) capture system-level regularities, but do not uniquely determine underlying causal mechanisms. Comparable macroscopic patterns could in principle arise from distinct micro-dynamics.

The observed dynamics result from the coupling of two feedback processes. Positive feedback operates through probability-weighted deposition, which increases local attractiveness in regions of existing accumulation. Negative feedback operates through stability monitoring: sites in which material persists for ten consecutive ticks increase in attractiveness (factor 1.2), whereas loss of material reduces attractiveness. Together these mechanisms produce selective consolidation. Configurations that withstand perturbation progressively bias subsequent construction.

Crucially, reinforcement is not blind accumulation but historically filtered persistence: only configurations that survive perturbation cycles continue to influence subsequent dynamics. Temporal asymmetry does not result from predictive modelling but from threshold-driven activation under stochastic resource distribution. Cross-cycle stabilisation follows from cumulative reinforcement of patches that survive perturbation. Retention is encoded in environmental attractiveness fields rather than agent-level recall.

The model does not introduce novel coordination rules at the local level; rather, it demonstrates how the interaction of probabilistic reinforcement and environmental memory can yield temporally structured regimes that are rarely isolated analytically in standard stigmergic analysis. Classic stigmergic models document emergent pattern formation (Deneubourg and Goss 1989; Bonabeau, Dorigo, and Theraulaz 1999; Camazine et al. 2003), but typically do not quantify whether such patterns exhibit forward temporal bias prior to disturbance, directional consolidation across perturbation cycles, or preferential reconstruction of historically stabilised configurations. By operationalising these three dimensions and tracking their co-occurrence, the present framework isolates a structurally constrained regime within feedback-mediated coordination.

Stigmergic accounts describe environmentally mediated coordination via positive and negative feedback (Deneubourg and Goss 1989; Bonabeau, Dorigo, and Theraulaz 1999; Camazine et al. 2003). The present model confirms that such mechanisms can generate temporal asymmetry and historically structured reconstruction. Whether this warrants a distinct conceptual category or represents a refined description within existing self-organisational theory remains open. The quasi-intentionality framework proposes that when stigmergic systems exhibit anticipatory bias, adaptive consolidation and historical reconstruction simulta-

neously, the resulting dynamics form a regime more constrained than purely reactive pattern formation. This claim requires comparative validation.

Enactivist and interactivist approaches argue that norm-sensitive organisation can arise without symbolic representation (Varela 1992; Thompson 2007; Barandiaran, Di Paolo, and Rohde 2009; Bickhard 2009). In biological minimal agency, normativity is grounded in viability conditions internal to the organism. In the present model, norm-sensitive behaviour derives from externally specified stability parameters and reinforcement rules. The comparison is therefore structural rather than ontological. Both domains exhibit distributed coordination shaped by constraint and history, but the grounding of normativity differs. The present analysis isolates structural preconditions of norm-sensitive coordination without resolving whether intrinsic normativity is constitutive of intentionality.

Teleonomic systems maintain functional parameters via feedback regulation (Mayr 1974). The present system differs in exhibiting selective reconstruction of historically stabilised configurations rather than simple restoration of equilibrium states. A thermostat restores a setpoint; here specific spatial patterns that proved stable under past perturbations are preferentially reinstated. The distinction is not one of complexity but of historical selectivity: regulation corrects present deviation, whereas reconstruction reinstates configurations whose persistence depends on prior survival under disturbance. Whether this difference marks a theoretically significant boundary or a graded extension of regulatory dynamics requires further analysis.

If quasi-intentionality is to function as more than a descriptive label, it must withstand comparative testing:

1. Ablation study
   Removing environmental attractiveness fields while preserving other parameters would test whether temporal asymmetry and retention depend on distributed memory. Collapse of reconstruction under ablation would indicate that trace-mediated coordination is constitutive of the observed regime.
2. Cross-domain implementation
   Applying identical probabilistic reinforcement rules to non-biological distributed systems would clarify whether the regime reflects general properties of feedback-coupled architectures or depends on specific spatial and dynamical constraints.
3. Biological validation
   Parameter calibration against empirical measurements from *Camponotus barbaricus* colonies would assess structural correspondence between model and biological construction dynamics.

The model demonstrates that anticipation, adaptation and retention can be operationalised as system-level properties in distributed architectures without central representation. These dynamics arise from probabilistic local rules coupled with environmental memory. The observed regime exhibits forward temporal bias, norm-sensitive restructuring under perturbation, and historically mediated reconstruction.

The central claim is deliberately restrained: restricting intentional vocabulary exclusively to representational systems leaves temporally structured, historically sensitive coordination without precise theoretical articulation. The present framework operationalises this domain

and specifies conditions under which it can be systematically investigated. Whether quasi-intentionality proves indispensable or reducible will depend on comparative and cross-domain analysis rather than definitional preference.

References

Barandiaran, Xabier E., Ezequiel Di Paolo, and Marieke Rohde. 2009. Defining Agency: Individuality, Normativity, Asymmetry, and Spatio-temporality in Action. *Adaptive Behavior* 17 (5): 367–386. https://doi.org/10.1177/1059712309343819.

Beer, Randall. 2014. Dynamical systems and embedded cognition. In *The Cambridge Handbook of Artificial Intelligence,* edited by Keith Frankish and William Ramsey, 128–145. Cambridge: Cambridge University Press.

Bickhard, Mark H. 2009. Interactivism: A manifesto. *New Ideas in Psychology* 27 (1): 85–95. https://doi.org/10.1016/j.newideapsych.2008.05.001.

Bonabeau, Eric, Marco Dorigo, and Guy Theraulaz. 1999. *Swarm intelligence: from natural to artificial systems.* New York: Oxford University Press.

Brentano, Franz. 2014. *Psychology from An Empirical Standpoint.* London: Routledge. https://doi.org/10.4324/9781315747446.

Camazine, Scott, Jean-Louis Deneubourg, Nigel R. Franks, James Sneyd, Guy Theraulaz, and Eric Bonabeau. 2003. *Self-Organization in Biological Systems.* Princeton: Princeton University Press. https://doi.org/10.2307/j.ctvzxx9tx.

Campbell, Donald T. 1974. "Downward Causation" in hierarchically organised biological systems. In *Studies in the philosophy of biology,* edited by Francisco J. Ayala and Theodosius Dobzhansky, 179–186. London: Palgrave.

Clark, Andy. 2016. *Surfing Uncertainty: Prediction, Action, and the Embodied Mind.* Oxford-New York: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780190217013.001.0001.

Clark, Andy, and David J. Chalmers. 1998. The Extended Mind. *Analysis* 58 (1): 7–19. https://doi.org/10.1093/analys/58.1.7.

Deneubourg, Jean-Louis, and Simon Goss. 1989. Collective patterns and decision-making. *Ethology Ecology & Evolution* 1 (4): 295–311. https://doi.org/10.1080/08927014.1989.9525500.

Dennett, Daniel. 1987. *The intentional stance.* Cambridge: MIT Press.

Dorigo, Marco, and Thomas Stützle. 2004. *Ant colony optimization.* Cambridge: MIT Press.

Emmeche, Claus. 1994. *The garden in the machine: the emerging science of artificial life.* Princeton: Princeton University Press.

Franks, Nigel R., Andy Wilby, Bryan W. Silverman, and Chris J. Tofts. 1992. Self-organizing nest construction in ants: sophisticated building by blind bulldozing. *Animal Behaviour* 44 (2): 357–375. https://doi.org/10.1016/0003-3472(92)90041-7.

Friston, Karl. 2010. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience* 11 (2): 127–138. https://doi.org/10.1038/nrn2787.

Haken, Hermann. 1977. *Synergetics: An Introduction Nonequilibrium Phase Transitions and Self-Organization in Physics, Chemistry and Biology.* Berlin; Heidelberg; NewYork: Springer-Verlag.

Hoeksema, Jason D., and Emilio M. Bruna. 2015. Context-dependent outcomes of mutualistic interactions. In *Mutualism,* edited by Judith L. Bronstein, 181–199. Oxford: Oxford University Press.

Hoffmeyer, Jesper. 2011. The Natural History of Intentionality. A Biosemiotic Approach. In *The Symbolic Species Evolved,* edited by Theresa Schilhab, Frederik Stjernfelt, and Terrence Deacon, 97–116. Series Title: Biosemiotics. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-007-2336-8_6.

Hopfield, John J. 1982. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences* 79 (8): 2554–2558. https://doi.org/10.1073/pnas.79.8.2554.

Kennedy, James, and Russell C. Eberhart. 1995. Particle swarm optimization. In *Proceedings of ICNN'95 - International Conference on Neural Networks,* 4:1942–1948. Perth, WA, Australia: IEEE. https://doi.org/10.1109/ICNN.1995.488968.

Kull, Kalevi, Terrence Deacon, Claus Emmeche, Jesper Hoffmeyer, and Frederik Stjernfelt. 2009. Theses on Biosemiotics: Prolegomena to a Theoretical Biology. *Biological Theory* 4 (2): 167–173. https://doi.org/10.1162/biot.2009.4.2.167.

Levin, Michael. 2019. The Computational Boundary of a "Self": Developmental Bioelectricity Drives Multicellularity and Scale-Free Cognition. *Frontiers in Psychology* 10:2688. https://doi.org/10.3389/fpsyg.2019.02688.

Lyon, Pamela. 2015. The cognitive cell: bacterial behavior reconsidered. *Frontiers in Microbiology* 6. https://doi.org/10.3389/fmicb.2015.00264.

Maturana, Humberto R., and Francisco J. Varela. 1980. *Autopoiesis and cognition: the realization of the living.* Dordrecht; Boston; London: D. Reidel Publishing Company.

Mayr, Ernst. 1974. Teleological and teleonomic: a new analysis. In *Methodological and historical essays in the natural and social sciences,* edited by Robert S. Cohen and Marx W. Wartofsky, 91–117. Boston: Springer.

Polak, Paweł, and Roman Krzanowski. 2023. How to Tame Artificial Intelligence? A Symbiotic Model for Beneficial AI. *Ethos. Kwartalnik Instytutu Jana Pawla II KUL* 36 (3): 92–106. https://doi.org/10.12887/36-2023-3-143-07.

Prigogine, Ilya, and Isabelle Stengers. 1984. *Order Out of Chaos: Man's New Dialogue with Nature.* Toronto; NewYork; London; Sydney: Bantam Books.

Searle, John. 1983. *Intentionality: An Essay in the Philosophy of Mind.* Cambridge: Cambridge University Press.

Thompson, Evan. 2007. *Mind in Life: Biology, Phenomenology, and the Sciences of Mind.* Cambridge: Harvard University Press.

Uexküll, Jakob von. 1921. *Umwelt und Innenwelt der Tiere.* Berlin: Julius Springer.

Uexküll, Jakob von. 1934. *Streifzüge durch die Umwelten von Tieren und Menschen: Ein Bilderbuch unsichtbarer Welten.* Hamburg: Rowohlt.

Varela, Francisco J. 1992. Autopoiesis and a biology of intentionality. In *Proceedings of the workshop on autopoiesis and perception,* edited by B. McMullin and N. Murphy, 4–14. Dublin: Dublin City University.

Wilensky, Uri. 1999. *Netlogo.* http://ccl.northwestern.edu/netlogo/ accessed March 1, 2026.

Ziemke, Tom. 2001. Are robots embodied? In *Proceedings of the first international workshop on epigenetic robotics: modelling cognitive development in robotic systems,* edited by Christian Balkenius, 75–83. Lund: Lund University Cognitive Studies.

Philosophical Problems in Science (2025), **79**, 73–90
(Zagadnienia Filozoficzne w Nauce)
doi:10.59203/zfn.79.730

ZFN Philosophical Problems in Science

ARTICLE

More than a tool: Artificial Intelligence in a philosophical history of technologies

Miroslav Vacura

Prague University of Economics and Business, Czech Republic
Email: vacuram@vse.cz

Abstract

This article examines artificial intelligence and shows how it is embedded in a history of technology. It develops a philosophical classification of tools as technologies based on the degree and nature of mediation that technology provides between humans and the world. Drawing on phenomenological and post-phenomenological conceptions—in particular the work of Ihde, Heidegger and Merleau-Ponty—it identifies five distinct stages in the historical development of technology, ranging from simple passive extensions of the body to advanced artificial intelligence systems. Each stage is defined by a qualitatively different mode of technological mediation between humans and the world, each transforming human perception and intentionality in increasingly complex ways. The most recent contemporary fifth stage in the evolution of technology, is characterized by the use of artificial neural networks, and the article demonstrates that a significant break from lower artificial intelligence systems separates this technology. Unlike simpler machines, these systems exhibit partial autonomy, are not epistemically transparent, and their outputs are unpredictable, forcing us to rethink their instrumental nature and the nature of instrumentality itself.

Keywords: postphenomenology, technological mediation, artificial intelligence, artificial neural networks

Introduction

One of the recurring themes of philosophical inquiry is the relationship between humans and technology, from ancient considerations of tools for crafts (τέχνη, *techné*) to contemporary debates about artificial intelligence and robotic systems (James 2019). This article considers artificial intelligence in the context of the historical development of tools and technologies, and proposes a systematic typology of technological development, based on phenomenological and post-phenomenological analyses and drawing on insights from cognitive science. Instead of the classification of technologies according to their social role and context, which is common among many authors (Mumford 1934; Marx 1990; Nolan and Lenski 2009), this study adopts a functional-structural approach based on emphasizing the nature of the mediation between human intentionality, technology and the world.

Taking as a starting point the works of Don Ihde, Martin Heidegger, Maurice Merleau-Ponty, James J. Gibson and others, our study explores the evolution of the human relationship

to tools, or technology in general, from simple hand tools to contemporary advanced artificial intelligence. Each stage of described historical trajectory shows the fundamentally different nature of tools or technologies, showing a progression from tools functioning as simple extensions of limbs to tools increasingly gaining independence and autonomy. Mapping this evolution not only allows us to better clarify the philosophical conception of tools and technologies, but also to provide a framework for understanding the ontological context of AI technologies.

The final part of this study focuses on advanced artificial intelligence systems based on artificial neural networks, showing that they represent a fundamentally new form of technology. These systems exhibit a high degree of autonomy, are epistemically largely unpredictable, and their behavior is nontransparent.

These systems go beyond the classical notion of tools, which understands them as an extension of the human body and intention. As a result, we show that it will be necessary to rethink a number of philosophical categories associated with the relationship between humans and technology. We conclude the text by speculating on a hypothetical sixth stage of technological evolution: the emergence of artificial consciousness.

Philosophical Approaches to Tools and Technological Mediation

The proposed typology of technologies and its conceptual framework is based on an exploration of several influential philosophical perspectives on tools and technological mediation. Drawing on phenomenology, postphenomenology and Gibson's ecological approach, this section delineates the essential components of the relationship between humans, tools and world. This analysis provides a starting point for understanding how technology transforms human perception, action and intentionality, and places in context the subsequent distinction between different stages of technological development.

If we wish to clarify the nature of the relationship between humans and technology, we must start with their distinction from a non-technological, natural, approach to the world. Ihde (1979, p.18) distinguishes two approaches to the world. The first is the non-technological approach to the world according to the scheme *Human* → *World*. The second is approach in which the instrument has a mediating role in the relationship between humans and the world: *Human—Instrument—World*. For Ihde, the simplest example of technology is the dentist's probe, a pointed piece of stainless steel that the dentist uses to probe his patient's teeth.

Other authors, e.g. Sullins (2011, p.152) in this context talk about the trinity of *user, tool and victim*—the term victim shifts the discussion to the moral domain, where we usually talk about situations where the user uses technology to injure a victim. This is, of course, a very specific approach that has its place in moral theory, but in this text we will draw on Ihde's approach, which is not so directly normatively loaded.

Ihde analyses the use of technology in general—a dentist's probe used to check the health of a patient's teeth is an informational tool—for it is used to collect information about the patient's teeth. In doing so, the dentist does not feel his probe with his hand but feels the tooth with the probe. Ihde's dentist is similar to the blind man, in the example given by Merleau-Ponty, who uses his stick to feel the space where he is walking—the stick is thus an extension of the blind man's touch: "The blind man's stick has ceased to be an object for him,

and is no longer perceived for itself; its point has become an area of sensitivity, extending the scope and active radius of touch, and providing a parallel to sight." (Merleau-Ponty 2005, p.165).

The realization of the intention to touch thus occurs at the end of the dentist's probe, at the end of the blind man's cane. Ihde (1979, p.19) says that the dentist feels the surface of the tooth at the end of the probe, not the probe at the point where it touches his palm—this is touch "at a distance". It turns out that not only sight and hearing are senses that can mediate sensory perceptions at a distance, but this is also the case for touch, which of course needs technological mediation. The technical enhancement of touch can take several forms based on different stages of technological development.

This type of tools is also thematized by Heidegger (2010, p.68) from which Ihde draws inspiration when he distinguishes different ways of human experience of the world. Human activity in the world largely consists of skillfully handling the things that are in the world. Useful things are in many cases simple tools—utensils for writing, utensils for sewing, utensils for working—but then there are often more complex technologies. These things have a pragmatic character, Heidegger refers to the Greek term πράγματα (*pragmata*), which refers to "things" in general, but is derived from the word πρᾶξις (*praxis*), denoting human activity or action. In this world view we thus understand things, as useful things, as tools [*Zeug*].

In this sense, it is possible that the Greeks, who understood things as πράγματα (*pragmata*), understood them primarily through their practical function, through their utility to themselves. We can consider in this context that they did not have a concept of a neutral thing that did not have a pragmatic purpose. A thing, the moment it becomes an object of human thought, always takes on a pragmatic aspect at the same time—at least in the background there is a consideration of what this thing can be useful for in human πρᾶξις (*praxis*) (Aristotle 1926, VI.2–5; Heidegger 2010, §15–18).

Heidegger emphasizes the relational character of these things—he says that a useful thing is essentially "something in order to [*um-zu*]..." (Heidegger 2010, §15), it thus involves a reference [*Verweisung*] to something else outside of itself. A useful thing is never available on its own, it is always embedded in a network of all useful things that only make it what it is: "There always belongs to the being of a useful thing a totality of useful things in which this useful thing can be what it is." (Heidegger 2010, §15).

Another related approach to this understanding of tools is the concept of affordances, developed by Gibson (2015) as part of his broader ecological approach, which was influenced by the work of Gestalt psychologists but had a number of original moments. The concept of affordances is based on the observation that when a person (or an animal) moves around in an environment, they do not perceive their surroundings homogeneously as a neutral space. The person perceives affordances on objects—places where an action is possible, places that the object "offers" to the agent, depending on the physical properties of both the agent and the object. Gibson argues that affordances are consciously perceived directly, they are not an afterthought derived from primarily perceived properties such as size, shape, and color. The agent directly perceives the possible points on which action is offered without any additional conscious processing of the perceived material.

A tool is an object that has affordances, offering us some "in order to" something else. We can generalize and extend the notion of a tool and consequently something else need not

be another object, it can be another state of affairs. The handle on the door offers itself to us as a "tool" whose use leads to a different state of affairs, one where the door is open.

Postphenomenological analyses also show how tools relate to society and the culture in which they are used. For example, Verbeek (2005, 2011) argues that technologies not only mediate action and perception, but also influence the moral space of human culture. Tools modify the experience of individual actions and influence which actions are considered morally acceptable and how responsibility is distributed within society.

Similarly, Rosenberger (2017) deals with technological mediations in the specific social contexts of their use. He emphasizes that the role of tools in society depends on the cultural environment and social practices that can influence how people learn to use tools and how they are subsequently integrated into human activities. Each tool allows for multiple uses, and specific practices are always based on social and cultural contexts.

A Functional Typology of Technological Development

In the following text we map the historical development of tools and technologies in general and divide it into six stages, which can also serve as a basis for the classification of current tools and technologies.

There have been a number of attempts in the history of philosophy to structure history in a similar way, but the distinction we propose below is based on the degree and nature of mediation that technology provides between humans and the world rather than other factors, and allows us to understand the breakthrough significance of AI technology.

For example, Comte (1830) in his famous distinction of the three stages was primarily focused on the stages of scientific knowledge, based on the way natural processes are explained and what role supernatural divine forces play in this explanation. For Marx (1990), the historical development of the social order, which was correlated with the modes of production, was of primary interest. Mumford (1934) distinguished the phases of the development of society according to the dominantly used materials and energy sources into eotechnic, paleotechnic and neotechnic phases. In contrast, McLuhan (2002) and similarly Ong (2012) divided technological development on the basis of the media and communication technologies used. Closer to our approach is the concept of Nolan and Lenski (2009), who distinguish phases of societal development by production techniques into Hunting and Gathering Societies, Horticultural and Pastoral Societies, Agrarian Societies, Industrial Societies and Post-industrial Societies. In their approach, the type of technology used already plays a more important role although here again it is significantly linked to the nature of the society. Similarly, Toffler (1990) distinguishes three waves of technological development: agricultural revolution, industrial revolution and post-industrial or information revolution.

In our distinction, we start strictly from technological leaps, which are based on an understanding of technology as a generalized form of human tool, and in this understanding of technology we draw primarily on the aforementioned post-phenomenological analyses of Ihde (1979), which draws primarily on Husserl and Heidegger.

Although these stages are presented in the following text as a historical sequence of development, they are rather historically emerged forms of technological mediation that can continue to coexist in contemporary society without excluding each other.

The order of the stages is not determined by moral considerations or any notions of ethical inferiority or superiority. Their order is based on when they appear in human history. The evaluative dimension contained in the typology relates exclusively to the type of technological mediation. Later stages surpass previous stages primarily in the sense that they further expand human capabilities, technological means gain greater autonomy, and they further transform the epistemic conditions of human access to the world and to these tools themselves. This hierarchy is therefore more historical-ontological and epistemic than axiological or ethical.

The First Stage

The earliest, first types of technological tools were purely static in nature, they were *extensions of the limbs* with possible enhancements of the sense of touch, such as just the stick, the mallet or hammer, or today the dentist's probe.

From a historical perspective, the first documented cases of this form of technological mediation can be identified in the late Paleolithic and early Neolithic periods. During these periods, the first stone tools appeared in archaeological finds—these were clubs and simple hammers. Examples include the Oldowan and Acheulean tool industries, in which tools function as direct extensions of the limbs, primarily improving strength or reach, without the use of more complex mechanical principles (Bower 1977).

However, this type of tool is not exclusively used by humans, but is also capable of being used by great apes. For example, Köhler (1976) has shown in his experiments with chimpanzees that they are able to use this type of tool. Human use of these early tools is distinguished by its stability within early communities and integration into their shared practices. Thus, it is only with humans that we can speak of the historically oldest permanent form of technological mediation.

Already Kapp (2015, p.40) understood these tools as "organ projections" (*Organprojektion*) of human faculties, i.e. for example the hammer as an enhancement and extension of the arm, optimized for a specific purpose. Heidegger (2010, p.69) calls such a tool "ready-to-hand", the hammer is ready to be put to work, no significant cognitive work is required. We do not approach the tool as a theoretical object with properties, we understand the tool through its use, through the activity we perform with it, that is, through πρᾶξις (*praxis*). When we use a hammer, we do not think about it, we take it in our hand and use it. The "ready-to-hand" tools are perceived in the activity as an extension of our body, they are directly integrated into the activity.

But we can approach the hammer in a second way, which Heidegger calls "present-at-hand"—in which case we stop, for example when the hammer breaks, and look at it theoretically as an object, rather than seeing it merely as a tool. In this stance, we try to understand the hammer through intellectual analysis and as such it is not available to animals. The essential point is that what makes an object a tool or technology is not only its properties, but also people's attitude toward that object.

The Second Stage

The second stage in the development of tools is represented by simple, non-powered mechanisms, such as a lever, sometimes consisting of several structural elements, such as a pulley.

Most are *simple machines* whose basic functional principle is that they *change the direction or magnitude of a force* based on simple physical principles. More complex mechanisms may involve, for example, interlocking gears. Their systematic use requires cognitive power that only humans have among animals. Most of these machines are not even capable of systematic use by apes. Chimpanzees, for example, are able to use a rod as a lever in some experimental settings, but this appears to be the result of random experiments rather than a systematic understanding of the principle of lever operation (Malherbe et al. 2024).

From a historical perspective, the emergence of this stage can be traced back to the beginning of the Neolithic period, to the development of the first agrarian communities in classical antiquity. At that time, simple machines were not only used in practice, but were also described theoretically in the form of early mechanical theories, e.g., in Greek engineering (Koloski-Ostrow 2008). Although these technologies do not use an engine, they require an abstract understanding of the mechanical principles of operation and application of force, which indicates a fundamental cognitive and cultural shift beyond technical mediation in the form of mere physical extension of the limbs.

The use of even the simplest mechanisms such as a lever is cognitively demanding. This is shown by the classic balance scale experiment proposed by Siegler (1976, p.481). In this experiment, children's understanding of proportional reasoning and their problem-solving strategies regarding the magnitude of force are studied. During the experiment, children are presented with a balance scale with different weights on either side and are asked to predict whether the scale will flatten or tilt to one side when released. More complex configurations of weights require more sophisticated reasoning strategies and the formulation of more complex rules. Siegler identified four stages of reasoning and rule complexity that children are capable of as they increase in age: stage 1 (approx. 4–5 years), stage 2 (approx. 6–7 years), stage 3 (approx. 8–9 years) and stage 4 (approx. 10+ years). This demonstrates that effective use of even simple machines has a non-negligible cognitive demand.

This experiment shows that reasoning regarding simple machines and other stage two tools requires significantly higher cognitive abilities than the first type of tools. A lever is a simple machine that operates in a similar way to Siegler's balance scale; it is a bar used in a specific way. This example also demonstrates that it is not just the tool itself that is important, but also its use—using the rod as a lever, using redirection of force, makes it a second stage tool.

At the same time, these are still often extensions of our limbs, especially in the case of simple machines like levers or pulleys. More complex mechanisms that require gears still often have to be driven by our own power, but their integration with our body is lower. However, they still have affordances and are seen as something in order to.

The Third Stage

The third stage of tool development comes when some *independent source of motion*, electric or otherwise, is available. While a pulley or lever can only increase the force by using the basic laws of physics, with the use of an additional source of power it is possible to increase the mediating effect far beyond simple machines.

Although the first machines using water power were already in use in ancient times (Wilson 2002), historically we can generally locate this stage in terms of its broader social

scope and impact at the beginning of the modern era and the Industrial Revolution (Ashton 1997). From the 18th century onwards, other external energy sources were systematically exploited, notably steam, followed by electricity and combustion engines. This transformation fundamentally changed the way and extent to which human activity affects the world. New types of machines no longer merely redirected human physical strength based on the laws of mechanics, but significantly amplified it, both in terms of its power and duration, using an external energy source. This change enabled forms of technological mediation that significantly exceeded the limits of natural human muscle power, yet remained entirely controlled by human will in every movement.

Examples include steam engines, locomotives, diesel-powered construction machinery, such as cranes, but also motorcycles or automobiles, for example. The word "automobile" is formed from the Greek word αὐτός (*autos*) meaning "self" and the Latin word *mobilis* meaning "mobile". They are therefore machines that have their own source of movement within them, self-propelled machines. However, they have no logic or will of their own; the self-source of motion is only intended to reinforce the intention and action of the human being. This is most evident in machines such as the construction crane, but it also applies to the traditional automobile (not modern automobiles equipped with sophisticated electronics or even artificial intelligence), as the latter only simplifies and accelerates human movement/running. In a traditional car, the gears are mechanical, the machine only supplies the extra force that sets the whole thing in motion. No inherent logic of the machine enters into controlling direction or speed.

Even in the case of machines like the crane, what was stated above in the case of the dentist and his probe applies. The centre of attention of the crane operator is the point where the object being carried is suspended. The experienced crane operator does not control the crane lever, but the crane arm and its suspension. The point where his hands touch the crane control levers is completely out of his attention, just as a walking man is unaware of the contraction of the thigh muscle. The crane is an extension of the crane operator's limbs, part of his extended self.

This is similar in the case of machines that enable us to move faster, such as automobiles. Merleau-Ponty (2005, p.165) gives the following example, "If I am in the habit of driving a car, I enter a narrow opening and see that I can «get through» without comparing the width of the opening with that of the wings, just as I go through a doorway without checking the width of the doorway against that of my body." (see also Grünbaum 1930).

The Fourth Stage

The fourth stage of instrument development consists of integrating the actual (usually electronic) logic into the instrument. The tool is no longer a mere mechanical mediator transmitting directly the user's intention, but can actively participate in the transformation mediating between man and the world.

These types of instruments have a character that the ancient Greeks referred to with the word αὐτόματον (*automaton*)—they act on the basis of their own will, a logic that is, however, very simple and straightforward. For example, Homer in *The Iliad* (Homer 2007, 5.749, 18.376) speaks of automatic gates or automatic movement of tripods intended for the gods.

Historically, however, this stage appeared much later, with the development of cybernetics in the mid-twentieth century and the development of analog and, in particular, digital computing technology. Automated machines using electronic control systems represent another form of technological mediation, where each movement is no longer controlled directly by human will, but where a human-independent control system is built directly into the machine itself.

Such a control system can take many forms, and these approaches are generally referred to as symbolic artificial intelligence. They include if-then rule systems and expert systems, logic-based knowledge representation methods (e.g., Prolog), formal ontologies, and a whole range of others (Russell and Norvig 2016). These systems can function independently of humans in the long term, learn, and process new information. Their activity is still fully determined by epistemically transparent deterministic operations in quantities that are comprehensible to humans and is therefore fully predictable and interpretable.

A modern example is an industrial robot, which works autonomously but does not require the constant human presence, instead having a fixed program that precisely defines each of its repetitive movements. The essential point is that the resulting action of the tool is determined by a clearly predictable output of the system, based on a clear, precisely defined sequence of rules, computational instructions, or straightforward connections of logic circuits.

In this case, technology becomes partially independent of humans—humans enter their goals into the machine in the form of instructions and settings, and the technology then works independently, performing preset operations and activities in a given order. The function of the machine is completely epistemically transparent; every movement and action has a clearly defined cause that can be identified at the level of a specific operation, which has a clear interpretation and behind which there is some intention of the human author. These systems are completely heteronomous, controlled by external input instructions, and have no autonomy. Such a machine can be very complex, with large sets of instructions and configuration data. For example, there are plans to create Robot Scientists that would be capable of automating all aspects of the scientific discovery process (Sparkes et al. 2010). However, the machine is still understood as a tool—something in order to—but affordances may be lacking, for example, in an industrial robot.

When we talk about independence from humans, we must distinguish between two forms of independence: operational/functional autonomy—the ability to act without direct human intervention; socio-epistemic autonomy—independence from human practices, cultural and social norms, and the resulting interpretive structures. In this case, we are referring to independence in the first sense. The concept of technological independence described in this stage of the analysis cannot be understood as separating the tool from human social contexts. It is operational autonomy in the narrow sense, meaning that the system can function over longer periods of time based on a set of rules and change its operations according to them without direct human supervision and active intervention. This form of autonomy is embedded in socio-cultural frameworks created by humans, particularly through decisions that are reflected in the very design of the sequence of rules according to which the machine operates.

Research in the field of the philosophy and sociology of science and technology (STS) shows that even very complex computing systems remain embedded in human socio-cultural practices and institutional structures. For example, Leonelli (2016) showed that both com-

putational systems and the data used in them are linked to a number of human practices (annotation, categorization, interpretation, data management) and are therefore not solely dependent on the functioning of the technology itself. However, recognizing this social embeddedness does not conflict with the fact of increased operational autonomy and independence from humans. On the contrary, it clarifies that operational autonomy at the level of machine functioning and its individual activities can be accompanied by increased interconnection with human practices at the socio-epistemic level.

The Fifth Stage

The fifth and currently highest level of tool development is represented by technologies equipped with advanced artificial intelligence based on artificial neural networks. These incorporate a high degree of autonomy and unpredictability. This is also a fundamental difference distinguishing this degree of technological mediation—it is epistemically opaque, meaning that the activity of these machines is not only independent, but also impossible to fully interpret and understand—these systems are often referred to as black boxes.

Historically, we can locate this stage at the turn of the 20th and 21st centuries, and it is closely linked to the development of deep neural network technology and machine learning architectures based on it. The very concept of artificial neural networks originated in the mid-20th century and developed more or less in parallel with symbolic approaches. Artificial neural networks (and connectionist models) are therefore neither historically nor conceptually newer than symbolic systems, but they were limited by theoretical and practical constraints. Theoretical limitations were gradually overcome from the 1980s onwards, when the mathematical foundations of modern neural networks were laid. Subsequently, advances in the availability of raw computing power and reductions in the cost of computer hardware made it possible to build large artificial neural networks of previously unimaginable complexity, with trillions of connections (parameters) between nodes, enabling the current development of artificial intelligence based on these approaches.

In traditional terminology, the term "artificial intelligence" includes both older symbolic (rule-based and similar) systems and new systems based on large artificial neural networks. According to our classification, symbolic systems belong to the fourth tool class, while advanced systems based on artificial neural networks belong to the fifth tool class. The use of the term artificial intelligence for such a wide range of systems largely obscures how significant shift in technology the introduction of large artificial neural networks represents.

Although, at a practical level, current artificial neural network-based artificial intelligence systems are programmed using standard deterministic programming languages and algorithms, yet, as a result of the vast quantity of artificial neurons and their interconnections (in the trillions) and learning capabilities, the outputs of these systems are often inexplicable in the standard way used in the case of symbolic artificial intelligence.

Artificial intelligence is usually divided into weak and strong (Russell and Norvig 2016, p.1020). The term Weak AI refers to systems focused on solving a single narrowly defined task, such as driving a vehicle, predicting the weather, or controlling the electricity supply. Conversely, the term Strong AI refers to hypothetical systems that solve complex tasks requiring the use of intelligence as well as or better than a human, but are also general-purpose, i.e., can solve the same breadth of task types as a human, without the need for

prior learning of individual narrowly defined task types. Strong artificial intelligence is also referred to as Artificial General Intelligence (AGI). Hypothetical, also not yet existing, artificial intelligence with significantly higher capabilities than humans is referred to as Artificial Superintelligence (Bostrom 2014).

Previously, the Turing test was considered the criterion for achieving advanced artificial intelligence (Turing 1950). The current most advanced artificial intelligence systems—large language models—have successfully passed this test (Biever 2023; Jones and Bergen 2025). However, if we classify large language models according to distinction of strong and weak intelligence, the consensus is that large language models are a very advanced form of weak AI that is focused on processing and producing high quality textual output. Large language models are based on probability theory. The first step in their use is to process extremely large sets of textual data from which probabilistic information about the relationships between words and expressions are extracted, and these are then encoded into the parameters of an artificial neural network. The resulting large language model can then be used to construct meaningful language outputs based on the input texts.

Thus, language models predict the likelihood of the occurrence of a subsequent linguistic expression on the basis of textual (context) in the range of up to several thousand words. An artificial neural network successful in this prediction typically has a range of several billion parameters.

The well-known application of large-scale language models is in popular chat applications, however, this technology has been applied in a large number of other applications. Chat programs (e.g., OpenAI ChatGPT) allow to generate coherent, meaningful and contextually relevant answers to the questions asked and to have a meaningful conversation. Large language models are based on probability theory (Brown et al. 2020), but they are able to synthesize texts in an innovative way that produces meaningful results. However, large language models do not understand language in the same sense that humans do. Some authors refer to them as stochastic parrots that merely repeat text with probabilistic word combinations without understanding its meaning (Bender et al. 2021).

Although the capability of large language models is based on probability the resulting text is not a mere rearrangement of the input data or a copy of it, but shows a high degree of originality and competence (Floridi 2023). The use of probability and artificial neural network technology does not diminish the capabilities of large language models, which is why other authors reject this comparison to a parrot (Arkoudas 2023). All these factors are the reason why large language models are usually classified as very advanced weak artificial intelligence.

At the same time, it should be emphasized that large language models lack the subjective conscious phenomenal experience, semantic grounding in experienced reality, and intentionality of the human kind that are typical of human thought and cognition and thus of the human way of being. Unlike humans, language models do not have a subjective "existence in the world"; they are not "at home" in the world (Pacovská 2024). The general consensus is that large language models do not have consciousness (Goertzel 2023; Vervoort, Mizyakov, and Ugleva 2023).

The question is to what extent can we consider large language models as tools. For example, Hongladarom and van der Vaeren (2024) say that ChatGPT goes beyond the classical understanding of the technological mediation of reality to the human, arguing that it

acts as a kind of hermeneutic agent. In this sense they claim that ChatGPT can to some extent "perform the human". Similarly, Thai Vo-Nhu (2025) rejects an instrumentalist approach to this technology, arguing that it cannot be considered just another neutral tool.

In this sense, large language models are transcending the capabilities of traditional technologies and reaching a new level of autonomy and independence from humans. While an industrial robot following precise rules can also work independently, its function is limited by predefined deterministic procedures that are precisely predicted and planned by humans. A traditional industrial robot is a tool that is several stages more advanced than a hammer, but its function is based on human planning. Man has pre-planned and thought through every aspect of its function in detail in his imagination, and every movement somehow accomplishes the purpose for which it was built and programmed. The traditional industrial robot and other similar technologies that belong to the fourth stage of development are merely tools working alone, their function fully subordinated to human goals, planning and imagination.

However, systems based on artificial neural networks go far beyond the functioning of rule-based systems or straightforward algorithms. Their behavior is neither fully predictable nor plannable. We can design and train an artificial neural network with a goal in mind, but the resulting behavior may always exhibit unexpected anomalies of deviation, sometimes harmful, but at other times highly original and beneficial. How it internally arrives at its decisions is opaque, often referred to as black-box systems or unexplainable artificial intelligence.

Despite their sometimes outstanding performance, current systems based on large neural networks have a number of shortcomings. For example, large language models are trained on large volumes of text data and therefore adopt and sometimes even amplify social biases contained in the training data (Bender et al. 2021; Weidinger et al. 2021). Large language models also tend to generate credible-looking but factually incorrect outputs—so-called hallucinations—in certain contexts, which stem from their reliance on probabilistic calculations rather than actual understanding of content (Marcus and Davis 2020; Ji et al. 2023).

Another problem is that how they internally arrive at their decisions is opaque, often referred to as black box systems or unexplainable artificial intelligence. The characterization of these systems as black boxes does not refer to a single phenomenon, but encompasses several dimensions of opacity. First, we are talking about technical opacity, which is due to the quantitative scope of the models, as they involve trillions of connections between nodes, making it impossible to directly control the model. Second, we are talking about epistemic opacity, where it is not possible to fully reconstruct the semantic causal paths leading from the input to the output, i.e., to explain and interpret why the neural network produced the output it did. Thirdly, we talk about phenomenological opacity—a neural network is not a transparent mediator of human-defined intentions, but works as an autonomous machine that generates outputs based on internal logic, which, although created in a human-defined learning process, is inaccessible to human interpretation.

These types of opacity have been comprehensively analyzed in the literature. Pasquale (2016) and O'Neil (2016) focus on the social risks of opacity. Zerilli et al. (2019) deal with epistemic opacity and the difference between explainability and understanding, and Rudin (2019) calls for the use of information systems whose decisions can be fully interpreted. At the same time, opacity motivates extensive research initiatives aimed at making the results

of these systems easier to interpret and explain—explainability is one of the main areas of research in artificial neural networks (Gilpin et al. 2018).

In general, epistemic opacity is considered a fundamentally new characteristic of these systems, which is one of the reasons for their inclusion in a separate stage of technological mediation development.

Can we still consider such epistemically opaque systems to be tools? Rather, we can compare the relationship between humans and modern artificial intelligence based on artificial neural networks to the relationship between humans and dogs (Sullins 2011, p.152). Humans have been breeding dogs for their purposes for thousands of years, and if by technology we mean the manipulation of nature to achieve human goals, then according to Sullins, domesticated dogs also fall under technology. Yet dogs have a certain degree of natural intelligence and we systematically train them to improve their abilities in tasks we define. Such a task can be both attacking another human and helping the visually impaired. Dogs' behavior is not always completely predictable, and it is not possible to train a dog to completely control and predict all aspects of its behavior. The dog is always a partially autonomous, individual being. Modern artificial intelligence systems based on artificial neural networks, or in the future robots that will incorporate such systems, are much more like dogs than older industrial robots. At the same time, similar to dogs in the past, robots based on artificial intelligence may act as assistants and guides for people in the future, which is why some authors demand that these technologies be generally available (Špecián and Špeciánová 2025). We therefore list them as a fifth separate stage of technological development in our classification to highlight their uniqueness and transformative nature with respect to society.

Are these artificial intelligence systems tools? In a sense, yes—but in the same way that dogs are tools, i.e., they are trained for our purposes and our goals, which they help us achieve. Unlike the lower-level tools, they are not just a tool, but have a degree of autonomy and independence beyond that, and in this they transcend the trait of tools.

Let us now return to Merleau-Ponty's conception of the tool quoted above. For the blind man, the cane is not an object, it is not perceived; the blind man feels directly with the tip of the cane, not at the point of contact of the skin of the hand with the cane. The tip of the cane is the sensitive area with which the blind person palpates the surface in front of him. It is an extension that extends the range and active range of the blind person's touch. Can a system based on artificial neural networks become a tool in this way? Can it become an extension of the natural abilities of man? In certain cases, yes, if it fully supports the human user—we are referring to various smart assistance systems that anticipate the user's intentions and correct their direct actions to match their intentions. However, in most of their applications, such as conversational language models, their autonomous nature will often be apparent and will at least partially remove the human user from their straightforward use.

The Sixth Stage

In this context, can we think about a hypothetical sixth, even higher level of technology? Above, we made a comparison between a dog and an advanced artificial intelligence system. However, there is one important difference that we have not mentioned—the dog has a certain level of consciousness, whereas the scientific consensus for artificial intelligence

systems based on artificial neural networks is that they do not possess consciousness (Goertzel 2023; Vervoort, Mizyakov, and Ugleva 2023).

Thus, as a sixth, hypothetical level of technology, we can consider a technology possessing artificial consciousness. It is important to stress again that we must distinguish the question of artificial intelligence from the question of artificial consciousness. Systems of strong artificial intelligence or even systems of artificial superintelligence may exceed humans in solving most tasks, but they may not have human-like consciousness.

We understand this stage as a borderline case, based on the conceptual structure of the proposed typology. The first five stages concern technologies that are mediating in nature—they are linked to human intentionality, but with each subsequent stage, this link loosens and the tools become ontologically independent, epistemically opaque, and transition from heteronomy to autonomy. However, these tools are still not subjects of experience; they remain unconscious technological mediators of human intentionality. If we consider the hypothetical possibility that technology could acquire artificial consciousness, we understand this as a way of exploring the extreme limits of the hierarchical scheme described above. A description from the perspective of purely instrumental mediation ceases to be adequate here, and categories associated with subjectivity are required to reflect the ethical issues associated with any entity possessing consciousness. mutual agency.

At this stage, we use the term "consciousness" in a phenomenal sense, which means that we understand it as a state in which a machine has subjective experience (Chalmers 1995) or a state of "what-it-is-like" (Nagel 1974). This is not just about the sophistication of a machine's functioning or the degree of its autonomy, but rather a hypothetical system that has conscious experience. This stage is currently speculative and is a borderline case, an ideal form, not a technological milestone. We do not prefer any of the competing theories of consciousness here—global workspace theories (Baars 1995; Dehaene 2014), higher-order theories (Rosenthal 2010), and integrated-information theory (Tononi 2008). However, the possibility of this stage assumes that consciousness is physically realizable in non-biological systems, which is not a claim accepted by all authors (Block 2009; Chalmers 2010). If it were not valid, this stage would never be realistically achievable through non-biological technologies.

The sixth stage, as a borderline case, also has a methodological role in this classification: it takes into account the difference between autonomy or unpredictability and consciousness. The fifth stage is characterized by a qualitative shift towards systems that are epistemically opaque, whose actions cannot be interpreted straightforwardly because they are based on learned patterns encoded in trillions of connections in an artificial neural network. These systems may outwardly show signs of a certain autonomy, i.e., they follow rules that they themselves have created in the process of learning from a huge amount of empirical data. This behavior may tempt some interpreters to assign agency or subjectivity in a strong sense—that is, to assign consciousness. By explicitly separating fifth-stage systems, capable of autonomous technological mediation of human intentionality, from hypothetical sixth-stage systems, capable of conscious subjectivity, we methodologically avoid confusion at this level.

Outline of the Proposed Stages

The stages described above are conceived as ideal types rather than historically accurate reconstructions of the sequence of scientific progress. The criteria distinguishing the stages are functional and epistemic, which means that some technologies may fall into borderline areas between stages. The stages are cumulative, meaning that technologies characteristic of higher stages often include attributes of lower stages. The following table clearly shows the characteristics of the stages described above.

Table 1. Stages in the development of technology

No.	Characteristics of the stage
1	Extension of limbs.
2	Simple mechanisms that change the direction or magnitude of a force based on simple physical principles.
3	Mechanisms using an independent source of motion.
4	Technology that integrates a control system, including rules, logic, or algorithms.
5	Technologies equipped with advanced artificial intelligence based on large artificial neural networks.
6	Technology with artificial consciousness.

The order in which the individual stages of technological mediation are presented reflects the increase in the scope and depth of mediation. The tools typical of the first stages of historical development directly expanded the capabilities of the human body, but they were still fully controlled by human will. Later technologies increasingly lead to the independence of tools from human will and independent activity, initially based on logic or rules, but later also on increasingly independent decision-making. This change also alters the structure of action and responsibility in ways that are ethically problematic in many cases.

This is also related to the trust with which we relate to the means of technological mediation, which is an implicit factor present in all stages of the proposed typology, a problem that has been explored by a number of authors (Lee and See 2004; Hoff and Bashir 2015). In the first forms of tool use, we can understand trust as embodied. In the case of a dentist's probe, we rely on it to faithfully transmit tactile information, so it is usable and works properly, and therefore can be trusted in the sense of a bodily extension. As a result of this implicit embodied trust, it recedes from thematic attention and is transparent in the sense explained by Ihde. Similarly, in stages two and three, trust is unquestionable, as tools and machines operate on causally epistemically transparent principles. This comprehensibility directly leads to trust, which, although lower than in the first case, is still stable. We trust a lever mechanism, a traditional mechanical machine, or a crane because their movements are predictable and understandable; in the case of incomprehensible activity, it is a malfunction that can be relatively easily diagnosed, understood, and eliminated based on causal principles.

In the case of fourth-generation technologies, which have internal logic, trust becomes more conditional. Once configured and launched, these systems can operate independently, without human intervention, but their symbolic, logical or rule-based nature ensures epistemic transparency. As in the previous case, in the event of unexpected or incomprehensible activity, it is possible to stop the machine, fully diagnose and understand the causes of the malfunction, and restore trust.

Technologies based on artificial neural networks falling within the fifth stage pose a fundamentally new problem of trust. These systems can no longer be trusted on the basis of predictability, epistemic transparency, and understanding of the causal chains of their internal functioning. Although these systems operate on causal principles, their quantitative scale (trillions of connections between nodes) makes it impossible to fully understand the causality of their actions, rendering them largely opaque to users. This and their autonomy undermine the basis on which simpler types of tools and machines were trusted as forms of extension of human intentionality. The problem of trust that arises from this situation is not only psychological, but also ontological. Such systems no longer function as technological means of mediation between people and the world, but achieve a degree of independence and autonomy as a result of which their actions can only be predicted probabilistically, and trust in them is thus considerably limited.

Conclusion

Drawing on phenomenological and postphenomenological frameworks, this paper assumes that tools and technologies are not merely neutral objects, but actively shape human perception, possibilities of action and corporeality. On this basis, we propose a typology of technologies based on the degree and nature by which technologies mediate between human intentionality and the world. This typology starts from simple hand tools and continues to contemporary artificial intelligence systems.

All stages of technological development—from early tools that exist only as extensions of limbs by means of rods, or to strengthen or refine them as in the case of levers and other mechanisms, to more complex machines that have their own power source and machines with built-in logic, to autonomous artificial intelligence systems—represent a fundamental advance in the relationship between tools and the human body.

The conclusion of our argument is that current artificial intelligence systems based on large artificial neural networks are not just an improvement on older, symbolic systems, but are a qualitatively new form of technology that is partially autonomous, fully unpredictable, and largely opaque, and thus to some extent uncontrollable by humans. In this sense, these AI systems go beyond the classical conception of tools as extensions of the human body or intentionality and force us to rethink existing conceptual categories.

At the same time, these systems do not yet possess consciousness and thus do not represent conscious agents comparable to humans. The hypothetical emergence of a technology equipped with artificial consciousness—if it ever occurs—would be another major ontological breakthrough, and we would understand such a technology as a sixth, fundamentally different type of technology, with its emergence leading to major philosophical implications.

Acknowledgement

The research reported in this paper has been supported by The Czech Science Foundation (GACR) grant no. 24-11697S.

References

Aristotle. 1926. *Nicomachean Ethics.* Translated by H. Rackham. Cambridge: Harvard University Press.

Arkoudas, Konstantine. 2023. ChatGPT is no Stochastic Parrot. But it also Claims that 1 is Greater than 1. *Philosophy & Technology* 36 (3): 54. https://doi.org/10.1007/s13347-023-00619-6.

Ashton, Thomas S. 1997. *The Industrial Revolution 1760-1830.* Oxford: Oxford University PressNew York, NY. https://doi.org/10.1093/oso/9780192892898.001.0001.

Baars, Bernard. 1995. *A Cognitive Theory of Consciousness.* Cambridge: Cambridge University Press.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency,* 610–623. Virtual Event Canada: ACM. https://doi.org/10.1145/3442188.3445922.

Biever, Celeste. 2023. ChatGPT broke the Turing test — the race is on for new ways to assess AI. *Nature* 619 (7971): 686–689. https://doi.org/10.1038/d41586-023-02361-7.

Block, Ned. 2009. Comparing the Major Theories of Consciousness. In *The Cognitive Neurosciences,* 4th ed., edited by Michael S. Gazzaniga. Cambridge: MIT Press. https://doi.org/10.7551/mitpress/8029.003.0099.

Bostrom, Nick. 2014. *Superintelligence: paths, dangers, strategies.* Oxford: Oxford University Press.

Bower, John R. F. 1977. Attributes of Oldowan and Lower Acheulean Tools: "Tradition" and Design in the Early Lower Paleolithic. *The South African Archaeological Bulletin* 32:113–126. https://doi.org/10.2307/3888658.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. 2020. *Language Models are Few-Shot Learners.* Version Number: 4. https://doi.org/10.48550/ARXIV.2005.14165.

Chalmers, David. 1995. Facing Up to the Problem of Consciousness. *Journal of Consciousness Studies* 2 (3): 200–219.

Chalmers, David. 2010. The Singularity: A Philosophical Analysis. *Journal of Consciousness Studies* 17 (9-10): 7–65.

Comte, Auguste. 1830. *Cours de Philosophie Positive.* Paris: Bachelier.

Dehaene, Stanislas. 2014. *Consciousness and the brain: deciphering how the brain codes our thoughts.* New York: Penguin Books.

Floridi, Luciano. 2023. AI as Agency Without Intelligence: on ChatGPT, Large Language Models, and Other Generative Models. *Philosophy & Technology* 36 (1): 15. https://doi.org/10.1007/s13347-023-00621-y.

Gibson, James J. 2015. *The Ecological Approach to Visual Perception: Classic Edition.* Hove: Psychology Press. https://doi.org/10.4324/9781315740218.

Gilpin, Leilani H., David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining Explanations: An Overview of Interpretability of Machine Learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA),* 80–89. Turin, Italy: IEEE. https://doi.org/10.1109/DSAA.2018.00018.

Goertzel, Ben. 2023. *Generative AI vs. AGI: The Cognitive Strengths and Weaknesses of Modern LLMs.* Version Number: 1. https://doi.org/10.48550/ARXIV.2309.10371.

Grünbaum, Abraham Anton. 1930. Aphasie und Motorik. *Zeitschrift für die gesamte Neurologie und Psychiatrie* 130 (1): 385–412.

Heidegger, Martin. 2010. *Being and time.* Translated by J. Stambaugh. Albany: State University of New York Press, SUNY Press.

Hoff, Kevin Anthony, and Masooda Bashir. 2015. Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 57 (3): 407–434. https://doi.org/10.1177/0018720814547570.

Homer. 2007. *The Iliad.* 2nd ed. Translated by I.C. Johnston. Arlington: Richer Resources Publications.

Hongladarom, Soraj, and Auriane Van Der Vaeren. 2024. ChatGPT, Postphenomenology, and the Human-Technology-Nature Relations. *Journal of Human-Technology Relations* 2. https://doi.org/10.59490/jhtr.2024.2.7386.

Ihde, Don. 1979. *Technics and praxis. Boston studies in the philosophy of science.* Dordrecht; Holland; Boston: D. Reidel Publishing Company.

James, Ian. 2019. Tekhne. In *Oxford Research Encyclopedia of Literature.* Oxford: Oxford University Press. https://doi.org/10.1093/acrefore/9780190201098.013.121.

Ji, Ziwei, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys* 55 (12): 1–38. https://doi.org/10.1145/3571730.

Jones, Cameron R., and Benjamin K. Bergen. 2025. *Large Language Models Pass the Turing Test.* Version Number: 1. https://doi.org/10.48550/ARXIV.2503.23674.

Kapp, Ernst. 2015. *Grundlinien einer Philosophie der Technik: zur Entstehungsgeschichte der Kultur aus neuen Gesichtspunkten.* Hamburg: F. Meiner.

Köhler, Wolfgang. 1976. *The mentality of apes.* New York: Liveright.

Koloski-Ostrow, Ann Olga. 2008. New Perspectives on Ancient Technology and Engineering in Greece and Rome. *Technology and Culture* 49 (4): 1025–1030.

Lee, John D., and K. A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 46 (1): 50–80. https://doi.org/10.1518/hfes.46.1.50_30392.

Leonelli, Sabina. 2016. *Data-Centric Biology: A Philosophical Study.* Chicago: University of Chicago Press.

Malherbe, Mathieu, Liran Samuni, Sonja J. Ebel, Kathrin S. Kopp, Catherine Crockford, and Roman M. Wittig. 2024. Protracted development of stick tool use skills extends into adulthood in wild western chimpanzees. Edited by Frans B. M. De Waal. *PLOS Biology* 22 (5): e3002609. https://doi.org/10.1371/journal.pbio.3002609.

Marcus, Gary, and Ernest Davis. 2020. GPT-3, Bloviator: OpenAI's language generator has no idea what it's talking about. *Technology Review* 294. https://www.technologyreview.com/2020/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-ai-opinion accessed January 9, 2026.

Marx, Karl. 1990. *Capital: a critique of political economy, Volume 1.* Edited by B. Fowkes. London; New York: Penguin Books in association with New Left Review.

McLuhan, Marshall. 2002. *Understanding media: the extensions of man.* Cambridge: MIT Press.

Merleau-Ponty, Maurice. 2005. *Phenomenology of perception.* Edited by C. Smith. London; New York: Routledge.

Mumford, Lewis. 1934. *Technics and Civilization.* London: Routledge & Kegan Paul PLC.

Nagel, Thomas. 1974. What Is It Like to Be a Bat? *The Philosophical Review* 83 (4): 435–450. https://doi.org/10.2307/2183914.

Vo-Nhu, Thai. 2025. *The Thief of Virtue: "AI slop" is more than just bad content.* https://blog.apaonline.org/2025/12/24/the-thief-of-virtue-ai-slop-is-more-than-just-bad-content accessed January 10, 2026.

Nolan, Patrick, and Gerhard Lenski. 2009. *Human societies: an introduction to macrosociology.* Boulder: Paradigm Publishers.

O'Neil, Cathy. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.* New York: Crown Publishing Group.

Ong, Walter J. 2012. *Orality and literacy: the technologizing of the word. 30th anniversary ed.* London: Routledge.

Pacovská, Kamila. 2024. Being at Home in the World. In *Platonism,* edited by Herbert Hrachovec and Jakub Mácha, 247–268. Berlin: De Gruyter. https://doi.org/10.1515/9783111386294-016.

Pasquale, Frank. 2016. *The black box society: the secret algorithms that control money and information.* Cambridge: Harvard University Press.

Rosenberger, Robert. 2017. *Callous Objects: Designs Against the Homeless.* Minneapolis: University of Minnesota Press.

Rosenthal, David. 2010. *Consciousness and Mind.* Oxford: Clarendon Press.

Rudin, Cynthia. 2019. *Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead.* Publisher: arXiv Version Number: 3. https://arxiv.org/abs/1811.10154 accessed March 14, 2026.

Russell, Stuart J., and Peter Norvig. 2016. *Artificial intelligence: a modern approach.* Pearson Education.

Siegler, Robert S. 1976. Three aspects of cognitive development. *Cognitive Psychology* 8 (4): 481–520. https://doi.org/10.1016/0010-0285(76)90016-5.

Sparkes, Andrew, Wayne Aubrey, Emma Byrne, Amanda Clare, Muhammed N Khan, Maria Liakata, Magdalena Markham, et al. 2010. Towards Robot Scientists for autonomous scientific discovery. *Automated Experimentation* 2 (1): 1. https://doi.org/10.1186/1759-4499-2-1.

Špecián, Petr, and Jitka Špeciánová. 2025. Universal Basic AI Access: Countering the Digital Divide. *Acta Informatica Pragensia* 14 (2): 272–281. https://doi.org/10.18267/j.aip.270.

Sullins, John P. 2011. When Is a Robot a Moral Agent? In *Machine ethics,* edited by M. Anderson and S.L. Anderson, 151–161. Cambridge: Cambridge University Press.

Toffler, Alvin. 1990. *The third wave.* New York: Bantam Books.

Tononi, Giulio. 2008. Consciousness as Integrated Information: a Provisional Manifesto. *The Biological Bulletin* 215 (3): 216–242. https://doi.org/10.2307/25470707.

Turing, Alan M. 1950. Computing Machinery and Intelligence. *Mind* 59 (236): 433–460.

Verbeek, Peter-Paul. 2005. *What things do: Philosophical reflections on technology, agency, and design.* Pennsylvania State University Press.

Verbeek, Peter-Paul. 2011. *Moralizing technology: understanding and designing the morality of things.* Chicago: The University of Chicago press.

Vervoort, Louis, Vitaliy Mizyakov, and Anastasia Ugleva. 2023. *A criterion for Artificial General Intelligence: hypothetic-deductive reasoning, tested on ChatGPT.* Version Number: 1. https://doi.org/10.48550/ARXIV.2308.02950.

Weidinger, Laura, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, et al. 2021. *Ethical and social risks of harm from Language Models.* Version Number: 1. https://doi.org/10.48550/ARXIV.2112.04359.

Wilson, Andrew. 2002. Machines, Power and the Ancient Economy. *Journal of Roman Studies* 92:1–32. https://doi.org/10.2307/3184857.

Zerilli, John, Alistair Knott, James Maclaurin, and Colin Gavaghan. 2019. Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard? *Philosophy & Technology* 32 (4): 661–683. https://doi.org/10.1007/s13347-018-0330-6.