# Representation of Concepts in AI: Towards a Teleological Explanation

Mattia FUMAGALLI [a] and Roberta FERRARIO [b]

[a] *Department of Information Engineering and Computer Science (DISI)*
*University of Trento, Italy*

[b] *Laboratory for Applied Ontology, ISTC-CNR, Trento, Italy*

**Abstract.** The representation of concepts is a lively research activity in several artificial intelligence (AI) areas, such as knowledge representation, machine learning, and natural language processing. So far, many solutions have been proposed adopting different assumptions about the nature of concepts. Each of these solutions has been developed for capturing some specific features and for supporting some specific (artificial) cognitive operations. This paper provides a *teleological explanation* of the most widely shared approaches in AI to the representation of concepts. The paper aims at providing four main contributions: *i)* an overview of the mainstream philosophical theories of concepts; *ii)* a categorization of a wide range of AI solutions inspired by such theories of concepts; *iii)* the proposal of a method for a comprehensive explanation of the current approaches to concepts in AI based on a *teleosemantic* perspective; *iv)* an illustration of how the proposed explanation could constitute a contribution in the context of explainable AI.

**Keywords.** Concepts, Information, Representation, Cognition, Semantics, Artificial Intelligence, Explainable AI.

## 1. Introduction

It is a widespread conviction, both in the psychological and in the philosophical literature [1,2], that concepts are to be taken as essential tools for human thought. While there is a lively debate on their specific nature, researchers are in total agreement on the pivotal role of concepts in adequately explaining many cognitive phenomena like *categorization*, *learning*, *induction*, *language understanding*, *planning*, *decision-making*, and so on [3].

Within the field of *Artificial Intelligence* (AI), there is a huge amount of work trying to enable most of the above mentioned cognitive phenomena in artificial agents. Consequently, when some cognitive capabilities need to be implemented into an artificial system, the tasks of choosing, modelling and organizing the best corresponding conceptual system have to be carried out. So far, many architectures have been realized adopting different approaches for the organization and the representation of artificial agents' conceptual systems [4]. All these computational architectures rely more or less explicitly on state of the art theories that provide an interpretation of the structure of concepts as tools for "thinking".

It can be generally observed that many different approaches to the computational representation of concepts have gone a long way with many success stories. Anyhow, by analyzing the current results in this area, two are the main considerations. Firstly, the criteria for evaluating the relevance[1] of a conceptual representation for enabling a certain cognitive (artificial) phenomenon are often implicit. Secondly, all AI approaches to the representation of concepts, considered in isolation, can efficiently account only for very few aspects of cognition. Some models, for instance, are used for enabling systems to reason on enormous amounts of data, but fail in accounting for trivial common-sense reasoning [6]. Similarly, some conceptual representations are impressively successful when used in well-defined domains, but they are completely inefficient in cross-domains settings [7].

This paper offers four main contributions. (i) Based on the evidence that the work in AI can take advantage of the philosophical research on concepts, we provide a brief overview of the mainstream theories on concepts. Our goal is not to provide a comprehensive survey of the state of the art theories on concepts. We refer the reader to excellent and thorough surveys, such as [3] or [8], for that purpose. Our central aim is, instead, to examine just some of the most relevant approaches, in order to make their assumptions explicit and link them to a common terminological (and theoretical) ground. (ii) We categorize a wide range of AI solutions on the basis of the introduced theories of concepts, shedding light on what is the task (e.g., classification, learning...) for which they are selected. (iii) Leveraging a *teleosemantic* perspective [9], we introduce a method for a comprehensive explanation of the current approaches to concepts in AI. This high-level theoretically grounded explanation may serve, in our view, as a blueprint for making explicit how a given approach to concepts may be relevant in relation to a certain artificial (cognitive) task to be addressed. (iv) Finally, we illustrate how the proposed approach may constitute a contribution in the context of explainable AI and may offer criteria for devising better solutions and, eventually, virtuous combinations of existing approaches.

## 2. Theories of Concepts: a Bird's-Eye View

In this section, we will try and sketch the main philosophical theories about concepts.

Among the most known theories of concepts, we can count the group of the so-called *classical-symbolic approaches*. According to these approaches, concepts are explicit representations codified in a language, similar to the first-order predicate calculus. The main features of this type of representations, also called *propositional representations* [10], are *arbitrariness* and *discreteness*. Concepts can be seen indeed as symbols of the *language of thought* (LOT) [11]. They are arbitrary in the sense that no similarity is needed between them and what they represent. They are discrete because they are either complex expressions separable in smaller parts, or atomic parts without any internal structure. Arbitrariness and discreteness allow the propositional representations of concepts to be highly formal, abstract and compositiitional.

Developed as an alternative paradigm with respect to the classical-symbolic approaches, the *connectionist research program* has a long story that dates back to the 40s [12], [13]. The many success stories of the symbolic approach around the 50s and 60s

[1] "*Something (A) is relevant to a task (T) if it increases the likelihood of accomplishing the goal (G), which is implied by T*" [5].

put connectionism in the shade for a long period. However, in the late 80s, it began to increase again its popularity. Connectionism shares the computational hypothesis of the symbolic approach, but provides a different model for concepts. In particular, according to this view, concepts can be seen as *representations distributed throughout a large number of processing elements*. Concepts are embedded in a *network* composed by *interconnected units*, which, at a certain level of abstraction, simulate the behavior of a conglomerate of neural cells [14].

A third family of approaches is grounded on the so-called *embodied/situated theories*, which hold that cognition is the product not only of what happens in the mind, but also in the body and in the environment [15]. The embodied theories program is quite recent and has not yet been consolidated in a wholly systematic theory, however it is being tested and used in many AI researches and applications (e.g., dynamical systems [16]).

We can distinguish also a group of less widely spread theories, which, however, play a pivotal role on the explanation of essential cognitive behaviors as well. Under this group we have *procedural theories*, being firstly asserted during the 70s and holding that it is not necessary for a concept to be explicitly represented as a mental symbol [17]. According to these theories, concepts can be implicitly represented as a procedure, i.e., as the execution of a piece of an algorithm. Within this framework, having a concept amounts to having a capability to do something. For instance, having the concept of ‘Cat’ is the same thing as having the ability of recognizing something as a ‘Cat’, or having the capability of using it in inference processes (i.e., inferring that it is an animal).

Moreover, another new interpretation of concepts was introduced around the late 60s, leading to the family of *analogical theories*, which purport that concepts are analogical representations (and not propositional, like in classical-symbolic theories). These kinds of representations are defined as mental objects that are similar to the objects they represent, like, for instance, a picture of a cat or the image of a cat on my eye’s retina [18]. Differently from propositional concepts, analogical concepts are not claimed to be discrete. This means that concepts do not provide a selection of features, rather they collect the whole perceptual information. This is a value if concreteness and completeness are the target, but it is a problem with respect to compositionality and abstraction [10]. Another interesting issue is that, with their representation of concepts, analogical theories provide an account for simulation processes in cognition, where specific distributed collections of information, captured through experience, may function as “collectors” of multiple conceptualizations for a single category (see for instance the notion of *proxytype* in [19] and the notion of *simulator* in [20]).

Finally, other three approaches to concepts that deserve to be considered here are: i) the *prototypical approach*; ii) the *exemplar approach* and the iii) *theory approach* [21,19]. Very briefly, in the prototypical approach, concepts provide the representation of the most “typical” occurrence for a given class of objects. Concepts are *prototypes*, i.e., a sort of weighted set of features (e.g., the prototype for ‘apple’ is something approximately round, green, red or yellow, with a specific range of weight, and so forth). In the exemplar view, concepts are seen as *devices* storing information about specific example occurrences for a given perceived object (e.g., the information about the apples we encountered in our experience). Within the “theory” approach, concepts are instead represented as *(micro-)theories*. For instance, having a concept for ‘apple’ means having a (micro-)theory about apples.

## 3. Theories of Concepts in AI

After this sketchy presentation of the main philosophical theories on concepts, we will now list in this section some AI approaches that have been inspired by such theories.

A typical example of a computational approach to concept representation that is influenced by classical-symbolic principles are *formal ontologies*. The most widely shared definition of ontology in the computer science community is Gruber's [22]: "a formal, explicit specification of a shared conceptualization"[2]. Ontologies can be seen as complex data structures, i.e., information artifacts, which can be designed (and formalized) using different representational languages, as for instance first order logic (FOL), or some computable fragment of it, like RDF[3] and OWL[4], following clear methodological principles (e.g., OntoClean [24]). All the languages used for representing these "conceptualizations" can thus be seen as instantiations of what in the classical-symbolic frame is taken as LOT. The main goal of these artifacts is to support *knowledge representation* (KR) and *integration* tasks, but they can be used for other tasks as well (for instance to drive NLP, or to provide data exchange formats).

*Neural networks* are typical computational representations inspired by the connectionist view of concepts. So far, even if they cannot be considered as proper models of real neural systems, different types of (artificial) neural networks have been successfully adopted for addressing specific AI tasks. These artifacts can be reduced to a set of interconnected units, i.e., abstract representation of neurons, where any connection between these neurons is an abstract representation of a synapse. According to these representations, each unit is associated to a numerical value, i.e., an activation state (or firing, namely the frequency by which a neuron sends signals through synapses). Each connection between neuron representation units is characterized by a weight that codifies the strength of that connection. The influence of a unit $x$ on a unit $y$ is given by the activation value of the unit $x$ multiplied by the weight of the connection from $x$ to $y$. The weight value can be positive or negative, so that the signal sent through the connection can activate or deactivate the neuron reached by the signal. So far, a lot of neural networks have been devised for capturing aspects of cognition, mainly connected with *learning*. For an overview and a collection of related papers we refer the reader to the "Neural Network Zoo" web page[5].

The AI approaches to concepts grounded on the embodied (or situated) theories are usually implemented by the *situated robotics* research program. A key exemplification of these approaches is the work by the research group at MIT[6], headed by Rodney Brooks [25]. This group is building robots that are equipped with simple sensory-motor devices and a collection of modules. Each of these modules is specialized for addressing a specific task, such as *checking for the presence* of an obstacle, *avoiding* an obstacle, *exploring*, and so forth. Each of these activities is run by a processor that works together with other processors and exchange information with the sensory-motor system and other processors. In these models no explicit representations are provided and no data is stored. The robots are not equipped with a mental model; they are automata that can be described

[2] A thorough analysis of this and other definitions of ontology may be found in [23].

[3] `https://www.w3.org/RDF/`

[4] `https://www.w3.org/OWL/`

[5] `http://www.asimovinstitute.org/neural-network-zoo/`

[6] `https://www.csail.mit.edu/`

just through finite states [26]. All the information used by these agents is grasped from the environment. Here concepts can be seen only as temporary representations, information flows, built upon the different phases of the perceptual process. The main goal is to derive the useful information from the environment, send it to the right processors and then produce an action. Thus, every robot can be seen just as a collection of behaviors in competition [27]. From an external point of view, it is possible to detect coherent behavioral patterns. However, locally, these robots are characterized by just casual processes. The robots devised following the situated approaches are able to reproduce the cognitive capabilities of some insects, and, according to recent results, it seems they can be evolved by introducing new connected processing modules.

In AI the *analogical* approaches are well-supported by research results like [18] and raise the issue of how some artificial cognitive process are related to *imagination* and deal with *mental images*. The underlying assumption of these computational frameworks is that perception and the relation with the external environment play a central role in cognition. This leads them to focus on the relevance of *simulation processes* and to share some hypothesis with the embodied approaches to representation. Though there are still few computational frameworks implementing the analogical approach, recently, some solutions grounded on this paradigm are being developed. For instance, the work in [18] aims at providing a computational account of cognition in modality-specific processing [2]. Examples of attempts at implementing simulations can be found in [28,29]. There are instead more computational frameworks implementing the ideas of procedural approaches. For those in AI starting from the procedural frame, the key idea is that concepts can be implicitly represented as *fragments of algorithms*. Concepts can be reduced to a sort of know-how that is not explicitly representable by means of data structures. However, these algorithms need some explicit information, or data structures, to work. Thus, procedural approaches do not exclude the possibility that the mental content is partially built on some explicit information, but they state that such content is mainly determined by the operations performed over it. Every representation is both involved in a causal relation with the external environment and in a causal relation with some mental operations. Good examples of computational frameworks linked to procedural semantics are *semantic networks* like KL-ONE [30] (for a detailed description see [31,32]) and *resources* like WordNet or FrameNet, inspired by Inferential Role Semantics (IRS), Lexical Semantics (LS) or Frame Semantics [33], i.e., semantic theories that underlie most of the procedural assumptions.

Finally, we can also find some works in AI explicitly developing some of the ideas grounding prototypical, exemplar and theory approaches. A computational work exploiting some of the features of prototypical and exemplar theories is the one by Lieto and Frixione [34], which is also partially inspired by the theory of *conceptual spaces* [35]. Here the main goal is to combine the typicality effects of a prototypical representation with the compositionality effects of a more classical representation of concepts [36]. The result is a sort of *hybrid architecture*, i.e., what they call DUAL-PECCS [37]. This is basically an integrated KR system aiming at supporting artificial cognitive capabilities, such as *categorization*, by implementing classical, prototypical and exemplar-based representations of concepts. For what concerns "theory" approaches, to some extent, we may say that *core ontologies* are examples of their computational applications. As an example, take the ORGANIZATION core ontology: as already shown, this is a typical formalism grounded on the symbolic frame, however it can be also seen as a (formal) micro-theory *representing the domain-specific concepts*.

## 4. Teleosemantics: a Teleological Approach to Concepts

In the following, we introduce a further philosophical theory on concepts, based on a view of concepts which is alternative to all the previous ones and that we would like to exploit for the development of a framework.

Teleosemantics provides an account of concepts and their representations by leveraging the notion of *(etiological) function* and the notions of *producer* and *consumer* devices[7]. According to this theory, the representations of concepts are kinds of *informational states* shared between a producer and a consumer device [9,38,39], which must be equipped with specific etiological functions. In the generation of a conceptual representation, the function of the producer is always to generate a state (the *representation*) according to a certain situation, i.e., when another state obtains (the *representatum*). The function of the consumer is to act in a certain way when the conceptual representation communicated by the producer has been received.

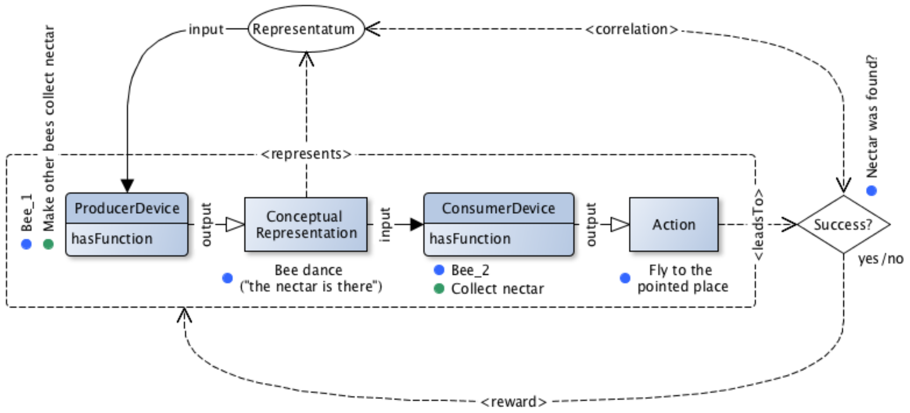

**Figure 1.** Overall view of teleosemantics core notions

Differently from more classical approaches, in which content of a conceptual representation is seen as its objective referent, according to teleosemantics, content is determined by a *success condition*, i.e., the condition that explains why a consumer acts successfully once a conceptual representation has been received. This situation can be easily illustrated by the example in Figure 1. Like many other animals, bees produce signals to inform other individuals of the same (or other) species. In Figure 1 we have a bee (producer device) that has the role of helping other bees in reaching a spot with nectar. The producer bee makes a special dance describing the distance and the direction to reach the place. The consumer bee (or some device within the bee) is responsible for

[7] Here the notion of device has to be taken in a broad sense, including, for instance, the perceptual apparatus of human beings and animals, or, according to a higher level of granularity, complex organisms.

the interpretation of the producer bee's dance. The conceptual representation is the bee's dance itself. In the example, the function of the producer is to generate a dance in certain circumstances and the function of the consumer is to interpret it and reach the nectar place. Since reaching the nectar place is only successful when the nectar is found (when there is no nectar the trip is only a waste of energy), this is the success condition of the actions/behavior prompted by the bee's dance. Consequently, the bee's dance means something like "the nectar is exactly there" [40].

Looking at the diagram shown in Figure 1, there is another important item to be explained, namely the one that is labeled as "reward" (see the edge from "Success?" to the group box), which represents what is obtained by the involved bees (devices) when their interaction is successful and that also accounts for the likelihood of this same strategy being adopted again in the future.

In these terms, describing what a conceptual representation represents amounts to considering the state of affairs that explains when the actions generated by the conceptual representation are successful. Such state of affairs is the condition represented by the conceptual representation.

The model represented above in Figure 1 can be then described more precisely as follows:

A conceptual representation *R* has content *x* iff:

- *R* can be defined as input/output in a system consisting of a producer (device) *P* and a consumer (device) *C*;
- The function of the producer *P* is to produce *R* to obtain *x*;
- *x* is the success condition, dependent on *R*, of *C*'s action caused by R.

The same schema illustrated in Figure 1 can be used to explain conceptual representations in artificial contexts and, more specifically, to classify AI approaches based on the function they carry out or, better, the function carried out by concepts in such approaches. Let us assume that the conceptual representation is in this case a data structure, like an ontology representing a classification of places and locations. In this situation, the producer can be defined as the knowledge engineer designing the ontology, who has the role of helping other agents in reasoning about geographical information. The consumer is responsible for the interpretation of that produced data structure. The conceptual representation is, as it has been already said, the data structure itself. In this example, the function of the knowledge engineer (producer) is to generate the ontology in such a way as to enable a correct navigation of the space and the function of the consumer agent (notice that here we assume that it may be an artificial agent as well) is to run effective inferences about geographical information and move according to what is inferred. Since running effective inferences about geographical information is only successful when the agent can move correctly in a specific area by following the instructions expressed or inferred from the data structure, this is the success condition of the actions/behavior prompted by the ontology.

Of course, this is only a very sketchy description of the teleosemantic framework. However, in the next section, we will try and illustrate how this framework can be applied to the case of Artificial Intelligence technologies, so as to classify them based on the function they are meant to accomplish.

## 5. A Teleological Representation of Different AI Approaches to the Representation of Concepts

In our view, the latter example shows that teleosemantics can be effectively used as a high-level model for explaining AI solutions based on specific theories of concept representation, and not only for explaining the representation of concepts in biological systems. The roles of producer and consumer devices may be played, indeed, by organizations, humans or even artificial devices (e.g., software applications or robots), which are able to produce and/or use a conceptual representation (just think about the automatic generation of ontologies or reasoners applied to them). It is not difficult to see each of these devices as equipped with the function in relation to which the conceptual representation is produced or consumed. The "Conceptual Representation" category (see Figure 1) captures AI approaches (theories and applications that operate on concepts, as internally defined) like the ones introduced in Section 3. "Action" can be seen as the category grouping kinds of executions of artificial cognitive tasks in which the conceptual representations can be involved. To have an idea of the items that could populate such category, just think about the group of AI activities introduced in [41], i.e.: *problem solving*, *knowledge* and *reasoning*, *acting logically*, *uncertain knowledge* and *reasoning*, *learning*, *communicating*, *perceiving* and *acting*. The "Success condition" class collects the descriptions of the "purposes" for which the "conceptual representations" are used by consumers. "Reward" represents what is gained once that the AI technology is successfully used.

The teleosemantics schema can also be used to characterize more specific AI technologies, not only kinds of approaches, as in the following example, where a specific data structure plays the role of "Conceptual representation":

- CONCEPT-REPRESENTATION: *"Google K-Graph"*
- PRODUCER: *"Google LLC"*
- PRODUCER-FUNCTION: *"to support people general information search"*
- CONSUMER: *"Person"*
- CONSUMER-FUNCTION: *"to find trustworthy medical information"*
- ACTION: *"to query Google search"*
- SUCCESS-CONDITION: *"trustworthy medical information is found"*
- REWARD: *"?"*

The above structure can be used to generate a database where different conceptual representations are classified and characterized according to their usage. This database can be a reference resource for people who need to assess, adopt and, eventually, combine existing AI solutions. Notice that the kinds of conceptual representations that can be put in the framework can be of different levels of granularity (and also that it can be applied either to types or tokens). For instance, in the above example, Google Knowledge Graph[8] could be easily replaced by the general notion of "Knowledge Graph".

What has been left undefined in the example is what the notion of "Reward" classifies. In the next section we will focus both on this specific notion and on a general framework to propose a novel perspective on *Explainable AI* [42,43].

[8] https://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html

## 6. A Telelogical Perspective on Explainable AI

Nowadays, society is experiencing a very rapid growth and increasing adoption of AI technologies. While it is quite obvious that AI is more and more able to enhance the quality of life of individuals and communities, researchers, practitioners, institutions and users should also acknowledge the fact that *new risks* could emerge from a more widespread use of such technologies. For this reason, at the beginning of April 2019, the European Commission (EC) released some ethics guidelines for *trustworthy AI*[9].

The document produced by the EC is aimed at ensuring adherence to European ethical principles and values and the implementation of AI applications, which are robust, not only from the technical point of view, but also from the social one, as these are seen as components of wider socio-technical systems.

In what follows we will try and show how the framework we are proposing may contribute to enable many of the requirements inspired by the ethical imperatives listed in the guidelines. Such contribution will be deployed in two main directions: *transparency* and *explainability*, which will then have an impact on further dimensions.

Transparency[10] will be ensured by the adoption of a framework that highlights and *makes explicit* all the aspects of an AI application which may create concerns. Less straightforwardly, explainability[11] will be granted by the *functional perspective* provided by the use of teleosemantics as a driving methodology which pinpoints the function AI applications play for the producer and consumer agents when the success conditions are met.

Transparency and explainability are referred to in the guidelines as *principle of explicability*, and this is what is explicitly addressed by the approach we are proposing. Nonetheless, indirect contributions are foreseen also for other principles: *human autonomy* could be enhanced by allowing human agents to decide to change the AI application in use and choose a better fit to their needs in case the success conditions are not met; *prevention of harm* could be improved by reducing the information asymmetries of the involved human agents; *fairness* could be favored by making human actors aware of the function the application should serve and of whether such function was successfully achieved, otherwise the use of the application may be contested. One thing that should also be noted is that the potential tensions between the realization of different principles can also be better singled out by making explicit the functions that the application is supposed to play for the different agents involved (producer and consumer).

As it should have become clear by now, the categories singled out by the framework we have produced seem especially fit to document the most important characteristics of an AI technology[12] from the point of view of its explainability: the conceptual representation (the data structure that underlies the technology), the producer (the agent who cre-

[9] `https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines`

[10] Transparency is defined in the guidelines as: “This requirement is closely linked with the *principle of explicability* and encompasses transparency of elements relevant to an AI system: the data, the system and the business models.”.

[11] Explainability is defined in the guidelines as: “Explainability concerns the ability to explain both the technical processes of an AI system and the related human decisions (e.g. application areas of a system). Technical explainability requires that the decisions made by an AI system can be understood and traced by human beings”.

[12] We use “technology” here in a very general sense, it may refer to a theoretical approach, and to an application as well.

ated or sold it to the consumer), the consumer (the agent who is going to use it), the function of the producer (what the producer wants the consumer to do, by the technology), the function of the consumer (what the consumer wants to accomplish with the technology), the action (that the consumer executes with the technology) and the success condition (the outcome of the action that, when successful, provides the conceptual representation with content) and, finally, the reward, what makes the producer and the consumer understand that the communication went through and can be re-used in successive, similar situations.

But there is another sense in which our approach targets explainability in AI, i.e. by characterizing ontologically the notion of *explanation*.

The term "explanation" may be used to indicate both a *process* and an *object* (which is a kind of *reification*[13] of the process, its outcome). As a process, explanation may be seen as a *social process*, a communication from the explainer (the producer) to the explainee (the consumer), which happens through a conceptual representation whose content is given by an action that shows that the communication was successful. The success condition of such communication is then reified in a reward, which is the *explanation as an object*.

We believe there is some similitude between this characterization we are purporting and the one proposed by Miller in [45], for two main reasons. The first is that Miller sees explanation as *post-hoc interpretability* and in fact we may say that a representation concept or a data structure is interpreted only once the consumer's successful action shows its content. In other words, content can be ascribed to representations (or AI technologies) only *ex post*, after that the consumer's actions have been successfully performed. Thus, explanation, even in our framework, comes out as a form of *abductive reasoning*. The second reason is that Miller sees explanation as connected to a weak notion of causality, a kind of *functional causality*, such that the explanation is not deterministically inferred by the cause, but is *selected* among many possible causes. This also means that both causality and explanation end up to be *contextual*. All this is very similar to the teleosemantic approach, in which, among the possible explanations that one could give, the one which shows that the representation (or the technology) worked (functioned) is selected.

Finally, a very nice aspect of the framework we are proposing is that it is adaptable to many scenarios in AI in which an explanation is required. This includes the examples of the previous section, in which the explanation is the description of the successful accomplishment of a specific task by a specific AI technology, but also our own account, in which the explanation is the description of the successful accomplishment of a kind of task by a kind of AI technology.

## 7. Conclusions and Future Work

The paper was aimed at providing a high-level model to support a comprehensive explanation of the current approaches to concepts representation in AI, based on the teleosemantics theory.

[13] We are intending here "reification" in a technical way, similarly as in [44], where reification is applied to roles and social concepts.

In order to be able to capture the main teleosemantics elements of such AI approaches, a preliminary step was to examine the most important ones at the state of the art, together with the philosophical theories on concepts by which they were inspired.

A further contribution of the paper is to show how the proposed teleological framework could constitute a theoretical tool to foster explainable AI.

Among the potential paths of research opened up by the current investigations, two we deem particularly worth to be pursued. A first perspective concerns the population of a database where existing AI approaches to concepts can be stored and characterized. We plan to devise the database schema through the definition of an ontology where the semantics of the main teleosemantic notions are made formally explicit. A second research direction concerns the central role of the notion of "reward". We wish to formalize reward within the teleosemantic explanation and try to understand whether, how, and in which case, this is related to the evaluation and usage of the stored/characterized approaches.

## References

[1] Lorraine K Tyler, HE Moss, MR Durrant-Peatfield, and JP Levy. Conceptual structure and the structure of concepts: A distributed account of category-specific deficits. *Brain and language*, 75(2):195–231, 2000.

[2] Lawrence W Barsalou. The human conceptual system. *The Cambridge handbook of psycholinguistics*, pages 239–258, 2012.

[3] Gregory Murphy. *The big book of concepts*. MIT press, 2004.

[4] David Vernon, Giorgio Metta, and Giulio Sandini. A survey of artificial cognitive systems: Implications for the autonomous development of mental capabilities in computational agents. *IEEE transactions on evolutionary computation*, 11(2):151–180, 2007.

[5] Birger Hjørland and Frank S Christensen. Work tasks and socio-cognitive relevance: A specific example. *Journal of the American Society for Information Science and Technology*, 53(11):960–965, 2002.

[6] Ernest Davis and Gary Marcus. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Commun. ACM*, 58(9):92–103, 2015.

[7] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484, 2016.

[8] D. Braddon-Mitchell and F. Jackson. *Philosophy of Mind and Cognition: An Introduction*. Wiley, 2006.

[9] Graham Macdonald, David Papineau, et al. *Teleosemantics*. Oxford University Press, 2006.

[10] Shaun Gallagher and Dan Zahavi. *The phenomenological mind*. Routledge, 2013.

[11] Jerry A Fodor. *LOT 2: The language of thought revisited*. Oxford University Press on Demand, 2008.

[12] Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943.

[13] Donald O Hebb. *The organization of behavior: A neuropsychological theory*. Psychology Press, 2005.

[14] Jerry A Fodor and Zenon W Pylyshyn. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71, 1988.

[15] F.J. Varela, E. Thompson, E. Rosch, and J. Kabat-Zinn. *The Embodied Mind: Cognitive Science and Human Experience*. The MIT Press. MIT Press, 2017.

[16] Randall D Beer. A dynamical systems perspective on agent-environment interaction. *Artificial intelligence*, 72(1-2):173–215, 1995.

[17] Philip N. Johnson-Laird. Procedural semantics. *Cognition*, 5(3):189–214, 1977.

[18] Roger N Shepard and Jacqueline Metzler. Mental rotation of three-dimensional objects. *Science*, 171(3972):701–703, 1971.

[19] Jesse J. Prinz. *Furnishing the Mind: Concepts and Their Perceptual Basis*. MIT Press, 2002.

[20] Lawrence W Barsalou, W Kyle Simmons, Aron K Barbey, and Christine D Wilson. Grounding conceptual knowledge in modality-specific systems. *Trends in cognitive sciences*, 7(2):84–91, 2003.

[21] Eleanor Rosch. Cognitive representations of semantic categories. *Journal of experimental psychology: General*, 104(3):192, 1975.

[22] Thomas R. Gruber. A translation approach to portable ontology specifications. *KNOWLEDGE ACQUISITION*, 5:199–220, 1993.

[23] Nicola Guarino, Daniel Oberle, and Steffen Staab. What is an ontology? In *Handbook on ontologies*, pages 1–17. Springer, 2009.

[24] Nicola Guarino and Christopher Welty. Evaluating ontological decisions with ontoclean. *Commun. ACM*, 45(2):61–65, February 2002.

[25] Rodney Allen Brooks. *Cambrian intelligence: The early history of the new AI*. MIT press, 1999.

[26] Amedeo Giorgi. The phenomenological mind: An introduction to philosophy of mind and cognitive science. *Journal of Phenomenological Psychology*, 40(1):107, 2009.

[27] Rodney A Brooks. Intelligence without representation. *Artificial intelligence*, 47(1-3):139–159, 1991.

[28] D Joyce, L Richards, Angelo Cangelosi, and Kenny R Coventry. On the foundations of perceptual symbol systems: Specifying embodied representations via connectionism. In *The logic of cognitive systems: Fifth International Conference on Cognitive Modeling*, pages 147–152, 2003.

[29] Randall C O'Reilly. Six principles for biologically based computational models of cortical cognition. *Trends in cognitive sciences*, 2(11):455–462, 1998.

[30] Ronald J Brachman and James G Schmolze. An overview of the kl-one knowledge representation system. In *Readings in artificial intelligence and databases*, pages 207–230. Elsevier, 1989.

[31] Nicholas V Findler. *Associative networks: Representation and use of knowledge by computers*. Academic Press, 2014.

[32] Ronald J Brachman. On the epistemological status of semantic networks. In *Associative networks*, pages 3–50. Elsevier, 1979.

[33] Jerry A Fodor. *Concepts: Where cognitive science went wrong*. Oxford University Press, 1998.

[34] Antonio Lieto, Andrea Minieri, Alberto Piana, and Daniele P Radicioni. A knowledge-based system for prototypical reasoning. *Connection Science*, 27(2):137–152, 2015.

[35] Peter Gärdenfors. *The geometry of meaning: Semantics based on conceptual spaces*. MIT Press, 2014.

[36] Antonio Lieto, Christian Lebiere, and Alessandro Oltramari. The knowledge level in cognitive architectures: Current limitations and possible developments. *Cognitive Systems Research*, 48:39–55, 2018.

[37] Antonio Lieto, Daniele Paolo Radicioni, and Valentina Rho. A common-sense conceptual categorization system integrating heterogeneous proxytypes and the dual process of reasoning. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[38] Fausto Giunchiglia and Mattia Fumagalli. Concepts as (recognition) abilities. In *FOIS*, pages 153–166, 2016.

[39] Mattia Fumagalli, Gabor Bella, and Fausto Giunchiglia. Towards understanding classification and identification. In *PRICAI*, 2019.

[40] Ruth Garrett Millikan. *On Clear and Confused Ideas: An Essay About Substance Concepts*. Cambridge University Press, 2005.

[41] Stuart J Russell and Peter Norvig. *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited,, 2016.

[42] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, 6:52138–52160, 2018.

[43] Filip Karlo Došilović, Mario Brčić, and Nikica Hlupić. Explainable artificial intelligence: A survey. In *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*, pages 0210–0215. IEEE, 2018.

[44] Claudio Masolo, Laure Vieu, Emanuele Bottazzi, Carola Catenacci, Roberta Ferrario, Aldo Gangemi, and Nicola Guarino. Social roles and their descriptions. In *6th Int. Conf. on the Principles of Knowledge Representation and Reasoning (KR-2004)*, pages 267–277, 2004.

[45] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019.

[46] Rudi Studer, V Richard Benjamins, and Dieter Fensel. Knowledge engineering: principles and methods. *Data & knowledge engineering*, 25(1-2):161–197, 1998.

[47] Marcello Frixione and Antonio Lieto. Representing concepts in artificial systems: a clash of requirements. *Proc. HCP*, pages 75–82, 2011.

[48] Aaron Sloman. How can we reduce the gulf between artificial and natural intelligence? In *AIC*, pages 1–13, 2014.