Computational Approaches to Concepts Representation: A Whirlwind Tour

Mattia Fumagalli[1], Riccardo Baratella[2], Marcello Frixione[2] and Daniele Porello[2]

[1]Free University of Bozen-Bolzan, Bozen-Bolzan, Italy.
[2]University of Genoa, Genoa, Italy.

Contributing authors: mattia.fumagalli@unibz.it; riccardo.baratella@edu.unige.it; frix@dist.unige.it; daniele.porello@unige.it;

Abstract

The modelling of *concepts*, besides involving disciplines like *philosophy of mind* and *psychology*, is a fundamental and lively research problem in several *artificial intelligence (AI)* areas, such as *knowledge representation*, *machine learning*, and natural *language processing*. In this scenario, the most prominent proposed solutions adopt different (often incompatible) assumptions about the nature of such a notion. Each of these solutions has been developed to capture some specific features of concepts and support some specific (artificial) cognitive operations. This paper critically reviews the most notable computational approaches to the representation of concepts. The main goals are: *i)* to provide a shared terminology for the desiderata of concepts and their computational representation; *ii)* to classify and assess the heterogeneous computational approaches according to the provided terminology; *iii)* to provide a reader who may not be very familiar with theories of concepts with an introduction to major themes in this research and with pointers to different research projects, and *iv)* to offer philosophers, and potentially AI practitioners, a well-informed guide for selecting among various (and possibly competing) computational representations of concepts.

Keywords: Concepts, concepts representation, knowledge representation, conceptual modelling, ontologies, machine learning, neural networks, artificial intelligence.

1 Introduction

The notion of "concept" (E. Margolis & Laurence, 2023) involves different disciplines (philosophy, psychology, and AI) that in one respect interact and have interacted successfully, but sometimes have different purposes and terminologies, which creates great confusion. The term itself can correspond to different things depending on the case. Over the years, the work in the modelling of concepts has led to the recognition that various types of conceptual representations are needed to account for certain classes of cognitive phenomena (Tyler, Moss, Durrant-Peatfield, & Levy, 2000; Jackson, 2021; Laurence & Margolis, 1999; Murphy, 2004).

Within the field of Artificial Intelligence (AI), many *Information Systems (ISs)* (Russell & Norvig, 2010) have been realized by adopting different approaches for the organization and the representation of their *conceptual system* (Vernon, Metta, & Sandini, 2007). The formalization of new tools, such as the *perceptual symbol system* approach (Barsalou, 1999) and the *proxytype theory* (Prinz, 2004), gathered from different theories of concepts, has been put forward in (Lieto, Minieri, Piana, & Radicioni, 2015) and (Pezzulo et al., 2013). Statistical approaches, such as neural nets, implementing dynamic and situated conceptual representations have been exploited (e.g., (McClelland & Rogers, 2003)). Computational approaches (e.g., simulation/embodied approaches) that ground conceptual information in modality-specific systems have been provided (e.g., (Roy, 2005)). It can be generally observed that, so far, all these different representations of concepts have gone a long way with many success stories. Anyhow, none of them can account for all aspects of concept representations and phenomena involving concepts.

Some models, for instance, are used for enabling systems to reason over enormous amounts of data, but fail in accounting for trivial common-sense reasoning (E. Davis & Marcus, 2015). Similarly, some conceptual representations are impressively successful when used in well-defined domains, but they are completely inefficient in cross-domain settings (Silver et al., 2016). Based on this evidence, the main consideration is that artificial systems can take advantage of all these different conceptual representations to address different tasks. Thus, the focus of modern AI on concepts and their representations makes the understanding of the notion of concept, and the knowledge of the core of conceptual theories, a key factor in this area of research. Making explicit the modelling assumptions behind the different approaches is, indeed, an important issue to be addressed whenever, for instance, a conceptual representation has to be devised and compared, or integrated, to other conceptual representations.

This paper explores key computational approaches to conceptual representation, with the following objectives:

(1) Offering an introductory overview of major themes in concept research and linking it to the research in AI.

(2) Suggesting a reference terminology for characterizing computational representation, which is informed by the philosophy of mind and psychology.
(3) Categorizing and assessing diverse computational approaches based on the reference terminology.
(4) Supporting discussions on the benefits and limitations of computational approaches to concepts, given the philosophical and psychological insights.

Research on the computational representation of concepts involves aspects of AI, psychology, and philosophy, to the point that it is sometimes difficult to label certain contributions as philosophical rather than psychological or concerning AI. At the same time, however, the large number of articles, the dispersion, and sometimes the fragmented nature of the literature make it difficult to have a clear overview. As a result, certain lines of research develop without considering other relevant contributions, risking rediscovering known results or pursuing directions that have already proven impractical. Hence the opportunity for a critical review that aims to offer philosophers (but not only them) an overview of the main approaches and of the major issues under discussion. In particular, computational research on concepts can offer philosophers insights that may prove relevant to their analyses. On the other hand, philosophers can identify problems for which they possess conceptual tools that may prove useful.

In addition, the proposed contribution may prove useful in areas such as conceptual engineering, which will benefit from categorization and assessment of the different kinds of conceptual representations for its goals of re-engineering old concepts and design new ones (see, for instance, works like (Isaac, Koch, & Nefdt, 2022) and (Chalmers, 2020)). Further, this work may also be relevant to projects such as the one by Hofweber, who has the goal of reaching substantial metaphysical conclusions merely by reflecting on our concepts, through the notion of *inescapable concept*. Indeed, throughout his inquiry, Hofweber never explicates what our representations of the world consist of, and if there are different ones, how these different kinds of representation support or not his project (Hofweber, 2024). Similarly, the content of this paper may be used to raise awareness among AI practitioners by informing them about key findings in the theory of concepts. This could lead to more informed decisions when selecting formalism-representation types, taking into account both the specific task and the intrinsic characteristics of the formalism. In this way, the contribution may assist in identifying similarities and differences between various approaches, which could be leveraged during the design phase rather than relying solely on task execution.

The paper is organized as follows: Section 2 groups the current approaches to concept representation into three main classes and provides their description, along with a brief overview of some remarkable computational implementations. Section 3 proposes a list of desiderata through which concept representations can be assessed and compared. Section 4 describes how the introduced desiderata are addressed by the classified approaches. Section 5

discusses what can be learned from the proposed analysis, and how it may be used for future work. Section 6 is a brief conclusion.

2 Three Broad Classes of Theories

Different theories about the nature of concepts have been proposed in psychology, neuroscience, and philosophy of mind and then implemented by specific AI approaches (see Figure 1 below). Most of these theories are grouped according to the literature into two main classes: *Good Old Fashioned Artificial Intelligence (GOFAI)* theories and *New Fangled Artificial Intelligence (NFAI)*, or post-classical theories.[1] Along with the GOFAI and NFAI theories, there are also theories combining assumptions that motivate both GOFAI and NFAI theories. From now on, we will call these theories *Complementary Fangled Artificial Intelligence (CFAI)* theories. These theories mostly originated within the so-called GOFAI theories, but over time they broke away from the original group, giving rise to a different category, addressing separate problems that are not covered within the GOFAI program.

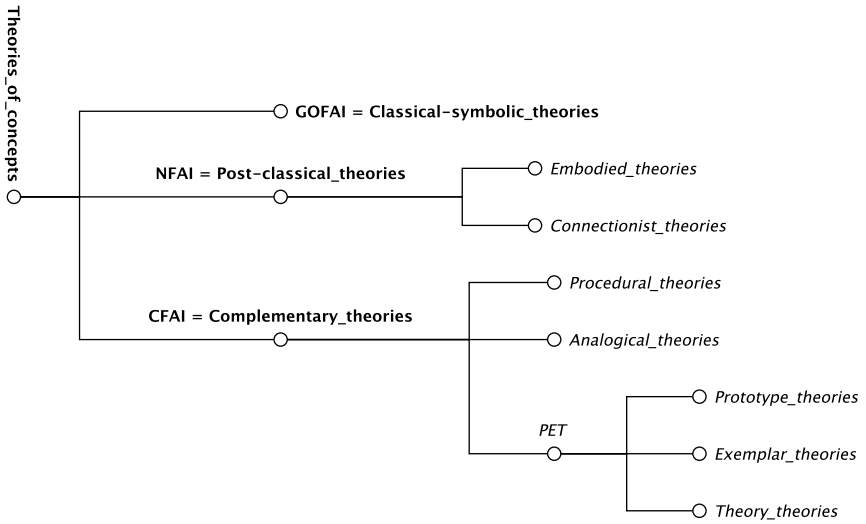


Fig. 1: Theories of concepts: a classification.

2.1 GOFAI

GOFAI theories, also known as classical-symbolic theories, present one of the most widely acknowledged perspectives on concepts. According to this view, concepts are explicit representations encoded in a logical language akin to

[1] Haugeland uses these umbrella terms in Haugeland (1989).

first-order predicate calculus (something similar to what J. Fodor termed a "Language of Thought" (LOT) in Fodor (2008)). These representations can exhibit varying levels of complexity and are typically characterized by a high degree of arbitrariness. That is, they are independent of the mechanism by which concepts may be acquired, e.g. involving the perception of some exemplars, and the form in which they are represented does not require to bear similarity with the actual mechanisms of classification.

GOFAI computational approaches. GOFAI computational approaches to a conceptual representation comprehend those that heavily rely on symbolic reasoning and manipulation of symbols to represent knowledge and solve problems. The main category under which those approaches can be classified is that of *Knowledge Representation and Reasoning (KRR)* (Levesque, 1986). Here a plethora of computational representations in the form of logical theories have been devised, to structure information, to enable rule-based reasoning, planning, and problem-solving.[2]

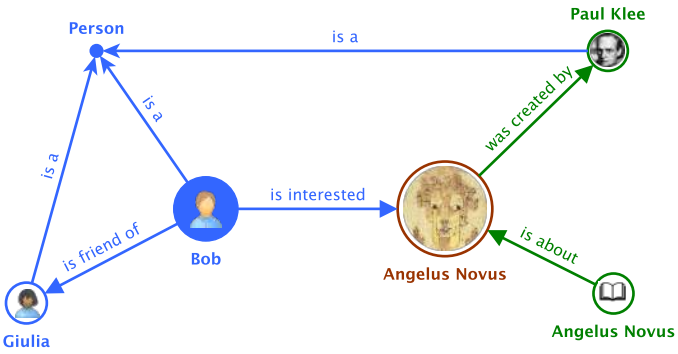


Fig. 2: A typical example of GOFAI concept representation.

A key example in this regard is the modelling of ontologies.[3] The most widely shared definition of ontology in the computer science community is *"a formal, explicit specification of a shared conceptualization"* (Guarino, Oberle, & Staab, 2009). In this definition, the term "formal" pertains to being machine-readable, "shared" indicates consensus within a group, and "conceptualization" is the process of defining an abstract model that describes certain knowledge. Ontologies in computer science can be broadly seen as logical

[2]To be historically accurate, GOFAI also made use of probabilistic frameworks such as Bayesian reasoning. However, we prefer to emphasize the symbolic aspect of GOFAI, which was the most prominent, and discuss probabilistic frameworks under NFAI and CFAI.

[3]Note that in the literature, especially from more recent years, we can find similar representations under the name of *Knowledge Graphs* (Chen, Jia, & Xiang, 2020) or *Vocabularies* (Vandenbussche, Atemezing, Poveda-Villalón, & Vatant, 2017). Still, these representations may have different characteristics, depending on the scenarios in which they are applied and the method by which they are constructed. For example, some knowledge graphs lack the distinction of concepts such as classes and instances, a key feature in ontologies.

theories written in a certain logical language. The languages that formally represent these "conceptualizations" may vary according to the expressivity or the computational complexity required for the specific modelling task. Prominent examples of computationally aware languages are *Descritpion Logics* (DLs), which provide a family of logical languages that correspond to decidable fragments of first-order logic, (Baader, Horrocks, Lutz, & Sattler, 2017). In fact, OWL (the *Web Ontology Language*), which provides the standard for representing concepts in the Semantic Web project, corresponds to a well-designed DL. By contrast, there are ontologies that require rich expressive logical languages. For instance, the main *top-level* or *foundational ontologies*, which are dedicated to modelling very general concepts (such as object, event, quality, space and time) and relations (such as parthood, constitution, temporal and spatial location), deploy first-order logic and quantified modal logic. Important examples of foundational ontologies are BFO (Basic Formal Ontology) (Otte, Beverley, & Ruttenberg, 2022), DOLCE (Descriptive Ontology for Linguistic and Cognitive Engineering) (Borgo et al., 2022), and (Borgo et al., 2022), UFO (Unified Foundational Ontology) (Guizzardi et al., 2022).[4] Additionally, ontology languages that go beyond first-order logic are discussed in the literature.[5]

Ontologies are often visualised utilizing graph data structures as the one represented by the fragment of Figure 2. Here we have nodes that may represent *classes* (e.g., "Person"), nodes representing *individuals* (e.g, "Paul Klee"),and *edges* representing different kinds of relations (e.g., "is about"). These graphs can be designed (and formalized) using different representational languages, see for instance RDF.[6]

Ontology languages can be seen as a perfect instantiation of what is taken as LOT in the classical-symbolic frame.[7]

The main goal of these artifacts is to support *information retrieval services* (e.g., search engines) and integration tasks (e.g., allowing to merge heterogeneous data), but they can be used for other tasks as well (see for instance driving *Natural Language Processing (NLP)*, or providing a data exchange format). To design well-behaved ontologies, several methodologies have been envisaged, most notably OntoClean (Guarino & Welty, 2002).

Foundational or top-level ontologies are dedicated to representing the most general concepts and relations required across most application domains, thus enabling cross-domain knowledge sharing. The top-level ontologies, such as BFO, DOLCE and UFO, are all designed according to well-defined principles and are used in many different ontology design and integration tasks. OWL-Time[8] and the ORGANIZATION[9] ontology are typical examples of what in

[4] Note that there are also OWL versions of BFO, UFO and DOLCE.

[5] See for example "Common Logic" an ISO standard for logic-based language (https://www.iso.org/standard/66249.html).

[6] https://www.w3.org/RDF/

[7] Here, ontologies are understood as computational artifacts with representational purposes, and we do not engage in the debate about the nature of what they represent, i.e., whether they endorse a realism-based approach, as discussed in B. Smith (2001).

[8] https://www.w3.org/TR/owl-time/

[9] https://www.w3.org/TR/vocab-org/

the ontology design community is called "core ontology", i.e., an ontology (more specific that top-level ontologies) expressing and specifying some concepts that can be shared among the different area of knowledge. OWL-Time is an ontology expressed in OWL-2,[10] describing temporal concepts, enabling the ability to express facts about topological relations among intervals and instants, together with information about the temporal position, frequency, and duration. The ORGANIZATION ontology provides a conceptualization representing the structure of organizations (e.g., business organizations, educational organizations, and so forth). It is designed to equip specific domain applications with information about organizations and roles. The WINE[11] ontology is another example of ontology, i.e., a domain ontology. Such an ontology is often used as a reference object for tutorials and ontology design tasks and provides a representation of wines, wineries, and all the objects needed for expressing this specific area of knowledge. SNOMED ontology is a domain ontology dedicated to biomedical concepts.[12]

2.2 NFAI

NFAI theories have been developed in recent years and are also known as *post-classical* theories (Haugeland, 1989). The NFAI class can be divided into two main sub-classes: the *embodied theories* and the *connectionist* theories. The embodied theories program still needs to be consolidated and cannot be considered as a genuine theory, however, it is being tested and used in many AI researches and applications (e.g., dynamical systems (Beer, 1995)). Differently from the embodied theories, the connectionist research program has a long story and dates back to the '40s (McCulloch & Pitts, 1943; Hebb, 2005). The many success stories of the symbolic approach around the '50s and '60s put connectionism in the shade for a long period. However, in the late '80s, it began to increase its popularity again. Connectionism shares the computational hypothesis of the symbolic approach but provides a different model for concepts. As discussed immediately below, concepts are embedded in a graph, where, at a certain level of abstraction, each node simulates the behaviour of a neural cell (Smolensky, 1988).

NFAI computational approaches. Neural networks are typical computational representations inspired by the connectionist view of concepts. So far, even if they cannot be considered proper models of real neural systems, different types of (artificial) neural networks have been successfully adopted for addressing specific AI tasks and supporting several applications (see for instance *Large Language Models* (Vaswani et al., 2017) like *ChatGPT*[13]).

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

[10] https://www.w3.org/TR/owl2-overview/
[11] https://www.w3.org/TR/owl-guide/wine.rdf
[12] https://www.snomed.org
[13] https://chat.openai.com/

unit is associated with 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 unit x on a unit y is given by the activation value of 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. Feed-forward networks (FF or FFNN) (Fine, 1999) are usually employed on pattern recognition tasks. These are powerful networks characterized by several layers of different units: *input units*, *hidden units*, and *output units*. The connections of FF networks are always unidirectional, i.e., they always start from an input unit through a hidden unit until an output unit. One interesting aspect of these networks is that they can be easily trained, i.e., they can learn how to produce results and tune their activation state by using a back-propagation mechanism. This allows the network to improve its reactions to given inputs and then improve its results. Besides FF networks we have many other (more or less recent) kinds of neural networks, for instance: *Recurrent Neural Networks(RNN)* (Medsker & Jain, 2001), *Radial Basis Function (RBF) Networks* (Broomhead & Lowe, 1988), *Hopfield Network (HN)* (Hopfield, 1982), *Markov Chains (MC)* or *Discrete Time Markov Chain (DTMC)* (Hayes et al., 2013), *Deep Belief Networks (DBN)* (Bengio, Lamblin, Popovici, & Larochelle, 2006) and *Deep Residual Networks (DRN)* (He, Zhang, Ren, & Sun, 2016). Each of them was devised to enable some specific artificial activities.[14]

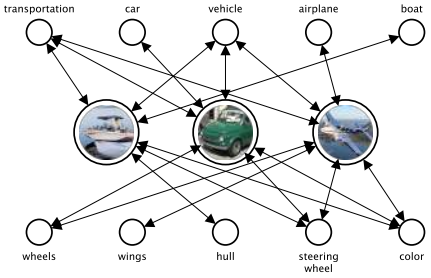


Fig. 3: An example of NFAI *connectionist* concept representation.

The network represented in Figure 3 is a typical example of RNN. The network represents certain information concerning certain vehicles (in the example we assume the training has already occurred and weights are already assigned to edges). The units in the centre are the so-called "hidden units", each one of them representing an instance of a vehicle. The bottom and top

[14] In this context, it is also worth citing approaches that discuss concepts as vectors, such as the one presented in Piantadosi et al. (2024).

units represent the features of these vehicles and can play the role of either input or output units. The hidden units are connected via bi-directional edges (with a positive value) with the features that characterize them. For instance, the instance of "boat" (far left) is connected to the feature "boat", "hull", "color", "steering-wheel", "vehicle" and "transportation".

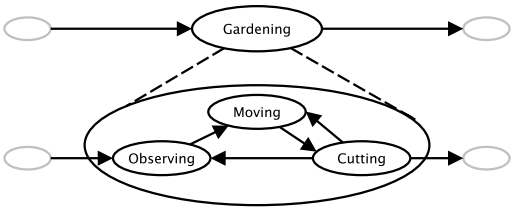


Fig. 4: An example of NFAI *embodied* representation.

Differently, the NFAI computational representations of concepts grounded on the embodied (or situated) approaches are usually implemented by the situated robotics research program. A key exemplification of these approaches is the work by the MIT research group, managed by Rodney Brooks (Brooks, 1999). 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 exchanges 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; rather, each activity is modelled by a finite state automaton (Gallagher & Zahavi, 2020). 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 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 entangled behaviours (Brooks, 1991). From an external point of view, in addressing certain tasks, it is possible to observe coherent behavioural patterns. For instance, the robots devised following the situated approaches seem to be able to reproduce the cognitive capabilities of some insects. As shown in Figure 4 (representation inspired by Craver and Bechtel (2007)) these patterns, which in the example may be represented by the ability to solve a simple task such as "gardening", are the result of a chain of processes or sub-activities (e.g., moving, observing, eye performing a saccade, cutting, etc.), where one may activate the other depending on the input received from the environment.

2.3 CFAI

CFAI theories are not in contrast with the GOFAI and NFAI programs. What we call CFAI theories rely, indeed, on assumptions that may be shared by both the previously described classes of theories. They can be seen as complementary views introduced to model aspects of cognition that are difficult to model with GOFAI and NFAI frames only. Under the category of CFAI theories, we group the *procedural theories*, the *analogical theories*, and the *prototype-exemplar-theory* (PET) theories. Procedural theories were raised during the '70s and their slogan says that a concept does not need to be explicitly represented as a mental symbol (Johnson-Laird, 1977). According to procedural theories, concepts can be implicitly represented as a "procedure", i.e., as the execution of a piece of an algorithm. According to this framework, having a concept is having the capability to do something. For instance, having the concept of 'Cat' is having the ability to recognize something as a 'Cat' or having the capability of using it in inference processes (e.g., classifying it as an animal). Similarly, the analogical theories, around the late '60s, introduced another new interpretation of concepts. According to these theories, concepts are analogical, namely, they are defined as mental objects that show structural similarity to the objects they represent, like, for instance, a picture of a cat or the image of a cat on my eye retina (Shepard & Metzler, 1971; Kosslyn, Thompson, & Ganis, 2006). Unlike representations in classical symbolic theories, which resemble elements of a language, concepts in these theories do not take on such linguistic resemblances. Note that here "analogical" is opposed to "propositional" and not to "digital": analogical representations in this sense are still digital representations. Another interesting issue is that, with their representation of concepts, analogical theories provide an account for simulation (see for instance proxy-types) processes in cognition (Prinz, 2004).

Regarding what we group under the PET category, we have three kinds of approaches to concept representation that face a similar problem, i.e., the *prototypical approach*, the *exemplar approach*, and the *theory-theory approach*. We can say that these programs were developed to account for concepts that exhibit prototypical effects. Following E. Rosch's work (Rosch, 1978), their main goal can be seen as offering a psychological model of concepts that explains category membership according to the criterion of *family resemblances* (Wittgenstein, 2009). According to these approaches, we do not represent a category by assuming its members share a set of common properties. Rather, among the members of the same category, there are more complex similarity relations, such as those between members of the same family. These approaches propose then solutions to justify our predictive capabilities, which do not seem to be well supported by the GOFAI program. According to the prototypical approach, concepts provide the representation of the "most typical" occurrences for a given object. Concepts are prototypes, i.e., a sort of weighted set of features (e.g., the prototype for 'Apple' is something round,

green, red, or yellow, with a specific range of weight, and so forth). In the exemplar view, concepts can be seen as devices storing information about specific example occurrences for a given object (e.g., the information about the apples we encountered in our experience). Within the theory-theory approaches concepts are represented as (micro-)theories. For instance, having a concept for 'Apple' means having a (micro-)theory about apples.

CFAI computational approaches. In AI the analogical approaches of the CFAI theories are well-supported by research results like the ones in Shepard and Metzler (1971) and Kosslyn et al. (2006), and raise the issue of how some artificial cognitive processes 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 share some hypotheses with the embodied approaches to representation. There is a lack of computational frameworks implementing the analogical approach. Still, some solutions grounded in this paradigm have been developed. For instance, see the work in Roy (2005), whose attempt is to provide a computational account of cognition in modality-specific processing (Gallagher & Zahavi, 2020). Examples of attempts in implementing simulations can be found in Cangelosi, Greco, and Harnad (2000), Cangelosi et al. (2005), Joyce, Richards, Cangelosi, and Coventry (2003), and O'Reilly (1998).

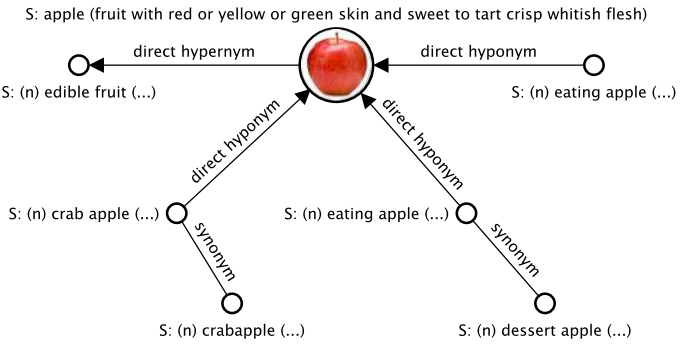


Fig. 5: An example of CFAI *(procedural)* concept representation taken from *WordNet*.

Differently from the analogical approaches, we have multiple 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 algorithms. Concepts can be reduced to a sort of know-how that is not explicitly representable when employing data structures. However, these algorithms need some explicit information, or data structures,

to work. The key point is that every representation encodes a sort of pragmatical knowledge, which is required to perform operations, which involve a causal relation with the external environment and a causal relation with some mental operations. Good examples of computational frameworks linked to procedural semantics are semantic networks and frames (Minsky, 1974), and resources like WordNet[15] or FrameNet[16] inspired by *Inferential Role Semantics (IRS)*, *Lexical Semantics (LS)* or *Frame Semantics* (Fodor, 1998), i.e., semantic theories that underlie a lot of the procedural assumptions. Figure 5 shows how the concept of "apple" is represented in WordNet. Here we have concepts that are represented by *synsyets*, namely terms (e.g., nouns or verbs) that are connected with other terms by a synonymity relation and with a corresponding gloss, plus hierarchical relations (e.g., *hyponymy* and *hypernymy*), denoting the semantic breadth of the concept. Moreover, we have results like the ones in Giunchiglia and Fumagalli (2016), Giunchiglia and Fumagalli (2017), or Fumagalli, Bella, and Giunchiglia (2019), with a particular focus on teleosemantics (Millikan, 2004), providing its (partial) formalization, its application in the context of KR, and its integration with the classical approach.

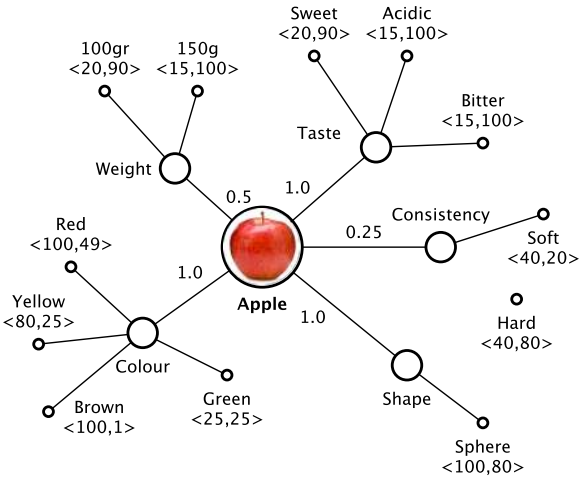


Fig. 6: An example of PET *(prototypical)* concept representation.

Among the complementary theories, PET approaches were particularly successful during the '70s, where in knowledge representation, with approaches such as *Quillian networks* and non-monotonic logics (Quillian, 1967), attempts were made to capture prototypical traits or "defeasible" aspects of concepts. Recently, some AI research has explicitly incorporated the foundational ideas

[15] https://wordnet.princeton.edu/
[16] https://framenet.icsi.berkeley.edu/

of these theories, trying to better cover some desiderata of their representations. An exemplar computational work exploiting some of the features of prototypical exemplar theories is the one provided by Lieto and Frixione (Lieto, Minieri, et al., 2015). This work is partially inspired by the theory of conceptual spaces (see Gardenfors (2014)). Such a representational approach is very well-known in artificial intelligence and proposes to model concepts as geometric structures that represent several quality desiderata, denoting basic features (e.g., *weight*, *colour*, *taste*, *temperature*, *pitch*) by which objects can be compared and categorized. 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. Still, the main motivation remains aligned with the one proposed by the prototype theory: each categorization task always consists of comparing new examples with a prototypical representation, that allows for checking similarities among objects. See for instance the prototypical representation of *apples* shown in Figure 6 (representation inspired by E. E. Smith, Osherson, Rips, and Keane (1988)). Here, each feature (e.g., "colour") is associated with a diagnostic value (e.g., 1.0), and each value (e.g., "red") is associated with a frequency threshold (e.g., from 100 to 49).[17] An exemplifying computational application in this regard is the hybrid architecture encoded by the DUAL-PECCS system (Lieto, Radicioni, & Rho, 2015). This is 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 the theory theories, to some extent, we may say that core ontologies are examples of computational applications. Just think of the organization's core ontology: as discussed in Sub-section 2.1, this is a typical formalism based on the symbolic frame, however, it can be also seen as a (formal) micro-theory representing the corresponding concept.[18]

3 Desiderata of Conceptual Representations

GOFAI, NFAI, and CFAI programs rely, more or less explicitly, on different assumptions about the nature of concepts and underlie different strategies for their representation. All these strategies of representation can be analysed by considering different desiderata. Within the context of this paper, by leveraging the analysis provided by the literature on "theories of concepts" (E. Margolis & Laurence, 2023; E. E. Margolis & Laurence, 1999; Sloman, 2014; Murphy, 2004; Braddon-Mitchell & Jackson, 1996; Gallagher & Zahavi, 2020; Kim, 2018), we selected 6 main desiderata. These are: *grounding*, *coverage*, *shareability*, *typicality*, *compositionality*, and, *incrementality*.

[17] Note that the numbers in the figures are arbitrary examples adapted from E. E. Smith et al. (1988).

[18] Please note that in the literature, alternative approaches exist that enable the definition of concepts approximately or less rigidly. These approaches often stem from classical methodologies such as description logic. See for instance the notion of *threshold concept* introduced in Baader and Gil (2024) and also discussed in (Porello et al., 2019).

The list does not claim to be exhaustive, but the selection criterion we have adopted is based on the level of recurrence of these desiderata in the literature about concepts we analysed (Gallagher & Zahavi, 2020). For example, the notion of compositionality is extensively addressed by several works (Fodor & Lepore, 2002; Millikan, 2004; Prinz, 2004; Rosch, 1975) and concerns a requirement considered fundamental to a theory of concepts. The same applies to the other desiderata.

A type of conceptual representation can be assessed w.r.t the way of addressing our list of desiderata. Let us look briefly at each of these in turn.

Grounding. This notion has been widely debated in the computational context (Harnad, 1990; Barsalou, 2008) and it is required for giving an account of how concepts are anchored in something perceptual mechanisms. Notice that, in this context, we use "grounding" as a synonym for "grounding in perception". In this sense, we may say that "grounding" is essential for explaining how conceptual representations are related to the environment they are about.[19] Within the various paradigms, an example of how the grounding desideratum can be addressed is the one provided by the so-called *causal approach to mental content*. According to this view, a concept of something in the world is a representation caused by this "something" (articulated in terms of sets of properties). The assumption here is that a concept C represents something S, if and only if S causes C (Adams & Aizawa, 2010). The basic idea is that any conceptual representation is derived by and covaries with the input perceptions of the encountered instances, according to a causal relation.[20]

Coverage. A desideratum of a conceptual model is that it can be used for representing all the types of concepts (Coliva, 2004; E. E. Margolis & Laurence, 1999) (see, for instance, *individual concepts* such as "Venus" or "José Saramago"; *relational concepts* such as "near"; *quality concepts* such as "yellow"; *living being concepts*, e.g., "animal" and "plant"; *stuff concepts*, e.g., "milk" or "gold"; *abstract concepts* such as "music" and "information"; *role concepts* such as "student" or "father"; *action concepts*, e.g., "creating" or "moving"). Within the different paradigms, we may have models providing an account for a large variety of concepts or models that are devised to provide an account for very few specific types of concepts only. For instance, some approaches mainly focus on how to represent concepts that are grounded in vision, like "physical objects", thus excluding those concepts concerning abstraction (e.g., "creative works" or "mathematical concepts").

[19] Note that model-theoretic semantics used in logic capture the compositional aspects of semantics, that is, how the meaning of complex expressions depends on the meaning of the components but they are silent on the semantics of primitive symbols. This is where the problem of grounding comes in.

[20] Consider that in the context of computational approaches, this desideratum is usually addressed when the provided formalism is devised to deal with data coming from sensors.

Shareability. Another important desideratum for concepts is the feature of being shareable (between humans and artificial agents). The aspect of shareability in concepts is usually facilitated by their explicit representation and pertains to the capacity of the involved agents to communicate the concept representation effectively. Across different paradigms, addressing the shareability aspect often involves representing concepts as particular types of symbols, namely descriptive constructs crafted by human designers. This is exemplified, for instance, by the association of "natural language labels" with nodes in graph data structures. The shareability desideratum poses several challenges, such as finding general standards for a common representation of concepts (Guarino et al., 2009), and in several approaches, it is a difficult constraint to satisfy. In fact, associating natural language with the adopted representations is often not easy, think of the case where concepts are modelled as activation patterns or as in the case of neural networks, as networks composed of several nodes. The opacity of approaches that lack a clear account of the shareability desideratum, poses a significant problem for conceptual representation reusability and explainability as well, as any of these representations can then be regarded as a sort of *black box*.

Typicality. Around the mid-'70s of the last century, the empirical results of Eleanor Rosch (Rosch, 1975) demonstrated the necessity of a new model for capturing both the structure of ordinary common sense concepts and the categorization processes. The results obtained by Rosch showed that most of the ordinary concepts often exhibit typical effects, i.e., they have common features that are central in the understanding and representation of the perceived objects and they can have some instances that are more typical than others.[21] Approaches addressing typicality are essential in improving tasks like categorization, recognition, or conceptual tracking (Frixione & Lieto, 2012) and can be used for enhancing other approaches focused on other desiderata.

Compositionality. This desideratum refers to the capability of producing infinite complex concepts starting from a finite set of atomic concepts. This is an essential feature for explaining conceptual systems productivity (Fodor & Lepore, 2002). Conceptual compositionality is well addressed by the approaches where concepts are considered as symbols and can be composed together through the application of certain *syntactic* rules that are mirrored in the *semantics* of the concepts. Mathematical logic is a perfect ally in this regard. Here the meaning of complex concepts can be determined by the meanings of their constituent concepts via the application of semantic rules governing their combination (Baader, Calvanese, McGuinness, Patel-Schneider, & Nardi, 2003). Differently, in some proposals, compositionality seems very difficult to address (Fodor & Pylyshyn, 1988) and this difficulty

[21] Note that this is the point that logic-based approaches fail to address. If two things are C, we cannot say what is the more typical.

leads to several problems, especially in enabling rational behaviours like reasoning and inference mechanisms, and explaining how the meaning of complex representations may depend on the parts by which they are composed.

Incrementality. The capacity to account for new information coming from the environment is a prerequisite for enabling the flexibility of biological cognitive agents and is one pivotal aspect of their cognition system (see, for instance, learning, evolution, and adaptation tasks). Some representations of concepts manage to capture this feature better than others. Usually, connectionist approaches are exploited in the context of applications where enabling a form of learning turns out to be the main feature. In this setting, see, for instance, the specific cases of *reinforcement learning techniques* (Kaelbling, Littman, & Moore, 1996) or *neuroevolution of augmenting topologies (NEAT)* (Papavasileiou, Cornelis, & Jansen, 2021). Here operations enabled by models can be considered *incremental* in the sense they can be updated from new inputs received from the context in which they are applied. Illustrative cases are the ones discussed in works like Kansky et al. (2017) or Olsen (2020), where these models can learn how to win video games improving their scores as they continue to play them. In principle, this desideratum can be also somehow accommodated by more symbolic approaches such as *default* (Besnard, 2013) *logics*, which, however, we do not consider here as typical examples of GOFAI theories. Still, these were found to be more focused on the management of incomplete information and less on the emergence of new functions and behaviours as in the cases of CFAI-based approaches, which, in this setting, are demonstrating outsized application success.

4 Assessing the Computational Approaches

In Figure 7, we list the computational approaches we have described, organizing them according to the classifications provided in Section 2 and the desiderata outlined in Section 3. Each approach is represented by a unique colour and then associated with a circle of varying sizes for each selected desideratum.

A *'small circle'* indicates that, given a reference approach, no well-established debate was found regarding how the desideratum is addressed. For example, we did not find any discussion on how neural networks address the shareability desideratum. Still, this does not imply that research about a specific approach does not account for the desideratum at all. There may be indeed sporadic cases where the desideratum is taken into consideration. However, these are cases that do not represent a well-established line of research with a clear literature debate associated with it.

A *'medium circle'* signifies that some works in the relevant literature explicitly address, discuss, or consider the desideratum, and there is a well-recognized debate in related work. This suggests potential connections between the desideratum and the approach, implying that the desideratum can be an

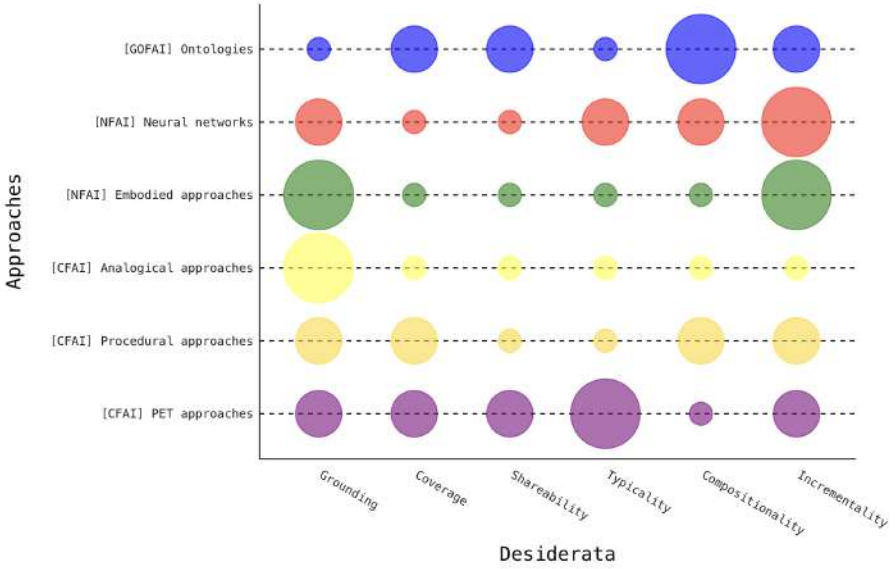


Fig. 7: Comparing different computational approaches: a summary overview. *'small circle'*: the target desideratum for the reference approach is not discussed or is exceptionally discussed in the reference literature. *'medium circle'*: the desideratum is part of a well-recognized debate in the reference literature. *'large circle'*: the desideratum is a key argument in the reference literature.

important indicator for evaluating the approach. Nonetheless, the desideratum is not always prominently discussed. For instance, while compositionality may not be captured by PET approaches, there is an ongoing debate on how prototypes can be compositional (Frixione & Lieto, 2012; Fodor & Lepore, 2002).

A *'large circle'* denotes that the desideratum is generally recognized as a fundamental feature of the approach. It is difficult, if not impossible, to consider the approach without acknowledging the desideratum. Moreover, the approach may have been specifically developed to address issues related to that desideratum. For instance, in ontologies, compositionality is a structural feature of the formalism used for their representation.

That being said, the proposed assessment is qualitative, based on a literature analysis.[22] Although not exhaustive due to the extensive time range and volume of work required, it can be used to indicate the level of connection between each approach and each desideratum.

Each decision can be then associated with a value on a scale: *'-1'* (*'small circle'*), *'0'* (*'medium circle'*), and *'1'* (*'large circle'*). These values correspond to:

[22] We adopted state-of-the-art manuals and articles and scraped *Google Scholar*, filtering articles by keywords related to approaches and desiderata, and number of citations. The extracted results, along with information and references from the manuals, were then filtered to obtain a consultation sample.

- *-1 = "not present or sporadic"* (we can find at most a few hints);
- *0 = "present"* (there is a well-recognized debate);
- *1 = "key"* (the desideratum is a pivotal aspect).

For instance, the literature widely supported and debated the fact that compositionally is a key aspect of GOFAI approaches. Differently, we can only find a niche debate to justify the relationship between GOFAI approaches and, for example, the desideratum of typicality.

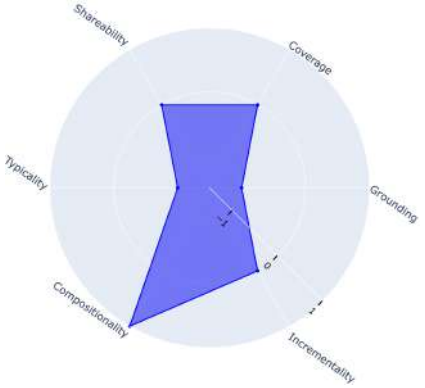
As a final remark, each assignment is flexible and subject to future updates. Future work could also refine this coarse-grained assessment into a more fine-grained evaluation. The highly-discrete values used here could be replaced with the number of works discussing the target desideratum for a given approach, with the circle size representing the number of articles found.

In the following, we provide a more detailed discussion and justification for the suggested assignments.

4.1 GOFAI

Desideratum	*State*	*References*
Grounding	**(-1)**	(Guarino et al., 2009), (Otte et al., 2022), (Ortmann & Daniel, 2011)
Coverage	**(0)**	(Siegel, Goolsbey, Kahlert, & Matthews, 2004), (Guha, Brickley, & Macbeth, 2016), (Lenat, 2022)
Shareability	**(0)**	(Gruber, 1993), (Studer, Benjamins, & Fensel, 1998), (Guarino et al., 2009), (Genesereth & Nilsson, 2012), (Vandenbussche et al., 2017)
Typicality	**(-1)**	(Frixione & Lieto, 2012)
Compositionality	**(1)**	(Fodor, 2008), (Baader et al., 2003), (Guarino et al., 2009)
Incrementality	**(0)**	(Adams & Aizawa, 2010), (Noy & Klein, 2004)

(a)



(b)

Fig. 8: A focus on the desiderata concerning ontologies: *(a)* reports the assessment value with some relevant references, *(b)* provides the corresponding visual representation.

Figure 8[23] provides information on the assessment of ontologies. Regarding the desideratum of **grounding**, we could not identify an established line of research addressing how to ground ontologies in perception. Although some ontologies are (or can be) explicitly developed under the premise of realism (Otte et al., 2022) and there are indeed articles investigating an empirical basis for ontological statements (Masolo, Botti Benevides, & Porello, 2018;

[23] We would like to remind the reader that the charts are meant to serve as a guide throughout the assessment, rather than as tools for quantifying the value of the respective approaches based on the selected desideratum.

Bottazzi, Ferrario, & Masolo, 2012), ontologies are generally not designed to process information from raw sensor data directly. Instead, these ontological modellings typically operate at a higher level of abstraction. Just looking at how ontology classes are usually characterized, we can find properties that can be used to identify the objects they categorize but have nothing to do with their perception. For example, classes such as "Person" or "Car", are characterized by properties such as "name", "model", or "ID". And even in case the ontology include properties such as colour or shape, they are not usually processed as low-level features extracted from sensors. For these reasons, we assigned **(-1)** in relation to this desideratum.

Considering **coverage**, we have a different situation. Ontologies are commonly developed to model a huge variety of concepts. As a check, we can take the huge ontological vocabularies collected in LOV[24] representing well-established projects aimed at covering as many concepts as possible. Let us take, for instance, *Schema.org*,[25] which can be expressed and formalized as an ontology (see its RDF formalization) and the set of its "commonly used types". We have concepts like 'CreativeWork', 'Artifact', 'Event', 'Organization', 'Person', and 'Place'; concepts like 'Action' (e.g., defining actions like 'Assess', 'Achieve', 'Move', 'Organize', along with 'Create', 'Reproduce', and so forth). Similarly, we have concepts of roles like 'Creator' or 'Student', and concepts of properties like 'Gender', 'JobTitle', 'Nationality', and so forth. However, the coverage desideratum cannot be considered as mandatory for ontologies. These artifacts can be devised just to cover a small portion of concepts (see all the projects related to the so-called domain ontologies, where we can find representations of a narrow domain of information, such as 'movies' or 'food'). Moreover, as anticipated for the grounding desideratum, it seems there are a few research works investigating the utilization of ontologies to portray concepts as entities directly rooted in perception, commonly referred to as "empirical concepts" (Bandrowski et al., 2016; Janowicz & Compton, 2010; Kuhn, 2009; Probst, 2008) (for further details, refer also to the differentiation between "nominal" and "empirical" categories as outlined in Giunchiglia and Fumagalli (2016)). This is the reason why we decided to assign **(0)** for this desideratum.

Works on ontologies also provide interesting insights on how to address the **shareability** desideratum. One of the scopes of ontologies is indeed *"to enable computers and people to work in cooperation"* (Berners-Lee, Hendler, & Lassila, 2001). In this sense, shareability is debated by well-established literature in this area. This is usually addressed by providing an explicit and formal representation of each concept. For instance, the concepts 'Person' and 'Nationality' map into a specific logical formula that can be, in principle, reused among different software agents and can be used by humans to understand the intended model behind the concept representations. However, we must emphasize how shareability is not always a primary goal of those

[24] http://lov.okfn.org/dataset/lov/
[25] http://schema.org/

developing ontologies. In fact, often, these artifacts are designed to support reasoning tasks without necessarily pandering to the constraint of interoperability and sharing with other standards. For this reason, we assigned **(0)** to the shareability desideratum.

The standard ontological formalisms do not address the desideratum of **typicality**, and this issue is generally not prioritized in this field (Yeung & Leung, 2006). Nonetheless, some researchers have made efforts to enhance ontologies to incorporate typicality. However, it cannot be said that there is a well-established research direction in this area. Although there are works that have addressed and continue to address this issue, such as those by Frixione and Lieto (2012) and Lieto and Pozzato (2018). Considering the previous point, although typicality is addressed in some research, it has not yet gained enough prominence to be widely recognized as a distinct research topic. Therefore, we are inclined to assign a rating of **(-1)** to ontologies.

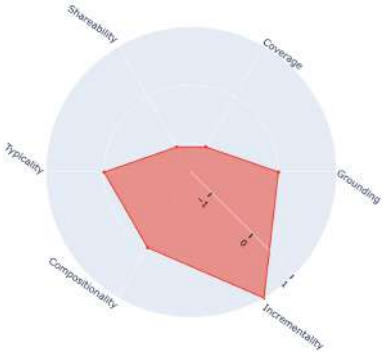
In contrast, **compositionality** is considered a key aspect for assessing the value of a given ontology, which is why we assigned a rating of **(1)** for this desideratum. A common agreement in the literature is that every ontology is a formal theory written in a certain logical language, with a clearly defined compositional syntax and semantics.

Finally, the default features of ontologies do not adequately address the **incrementality** desideratum as we described it. However, there is substantial and well-established research on ontology evolution and adaptation (see (Noy & Klein, 2004) for more details and (Besnard, 2013; Klir & Yuan, 1995) for examples of formalism), and incrementality is considered an important aspect of some hybrid approaches designed to enhance standard classical models. This scenario led us to assign a value of **(0)** for this desideratum.

4.2 NFAI

Desideratum	*State*	*References*
Grounding	**(0)**	(Rosenblatt, 1958)
Coverage	**(-1)**	(Fodor & Pylyshyn, 1988), (Hupkes, Dankers, Mul, & Bruni, 2020)
Shareability	**(-1)**	(Smolensky, 1988)
Typicality	**(0)**	(Kim, 2018), (Way, 1997), (Berthold, Sudweeks, Newton, & Coyne, 1998), (T. Davis & Poldrack, 2014)
Compositionality	**(0)**	(Fodor & Pylyshyn, 1988)
Incrementality	**(1)**	(Smolensky, 1988)

(a)



(b)

Fig. 9: A focus on the desiderata concerning neural networks.

Figure 9 provides information on the assessment of neural networks. Concerning **grounding**, connectionist representations of concepts share the same goal as classical views. Both approaches aim to model the mind's ability to be about something and are involved in tasks that simulate *intentional activities*. However, while classical approaches focus more on compositionality and shareability, making them less tied to the notion of "grounding" (i.e., focusing on "how" representations are created rather than "where" they are derived from), connectionist approaches (e.g., neural networks) address this desideratum more comprehensively. It is no coincidence that one of the applications of connectionist models is object recognition. Neural networks are often employed to infer new information from given inputs, enabling induction processes that discover unknown properties of objects based on known properties (e.g., *Feedforward Neural Networks (FFNNs)*). Still, we found no evidence in the literature indicating that this is the primary purpose of these formalisms. Moreover, the application of neural networks extends far beyond areas such as vision, addressing contexts devoid of perceptual data (e.g., financial or time series predictions). For this reason, we assigned a rating of **(0)** for this desideratum.

In contrast, **coverage** does not appear to be a priority within this framework. As discussed, the concepts represented by this approach are often "empirical",[26] and there seems to be no evidence of work aiming to cover a wide range of concepts. Additionally, models based on neural networks are recognized as weak in their ability to generalize and transfer knowledge between different domains. This limitation is certainly related to a poor connection with the coverage desideratum. For this reason, we assigned a rating of **(-1)** to this desideratum.

A similar situation exists for the **shareability** desideratum. Deriving a shareable representation of the content produced by these artifacts is rather difficult. In this context, concepts are not explicitly represented but are derived from certain properties of the network. This limits their human understandability and the possibility of sharing them across different situations and agents. Furthermore, there are no significant references to this desideratum in the relevant literature. While some literature on explanation in AI considers this desideratum relevant (Byrne, 2023; Guizzardi & Guarino, 2024), it is difficult to claim that this represents a well-established line of research. Therefore, we assigned a rating of **(-1)** to this desideratum.

By contrast, there is a well-established line of research discussing how neural networks can address the **typicality** desideratum, even if this cannot be considered a key characteristic of this formalism. These information artifacts can be naturally used to generate prototypical representations by generalizing from the collected data (Kim, 2018; Way, 1997; Berthold et al., 1998; T. Davis & Poldrack, 2014). Still, we cannot claim that typicality is always taken into

[26] Most of the models within this framework are aimed at recognizing perceptual information. Still, there are models used to predict more abstract concepts (see, for instance, time series, weather forecasting, etc. In this sense we can say that neural networks allow for the representation of different kinds of concepts).

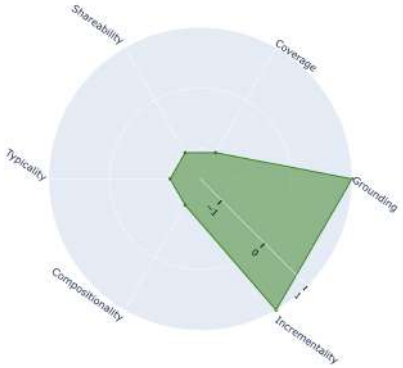
account in every work on neural networks. For this reason, we assigned **(0)** to this desideratum.

Similarly, **compositionality** does not appear to be a priority in the research agenda for these models since they do not commit to formal semantics (Santoro et al., 2017). Jerry Fodor and Zenon Pylyshyn argue that neural networks (NNs) and connectionist models are completely at odds with the requirement of compositionality (Fodor & Pylyshyn, 1988). However, even though NNs are not designed for logic-based systems, they are devised according to well-defined formalisms. There are works discussing how they can be mapped to logic, such as the hybrid approaches described in (Sun & Alexandre, 2013) and (Fodor & Pylyshyn, 1988). For this reason, we assigned **(0)** for this desideratum.

On the contrary, for what concerns **incrementality**, we may say that this is a core desideratum of the connectionist approach. Neural networks can be indeed the result of a training process only because of the adaptation and evolution capability of the entire net. Consequently, we assigned **(1)** to this desideratum.

Desideratum	*State*	*References*
Grounding	**(1)**	(Brooks, 1991), (Brooks, 1999), (Barsalou, 2008)
Coverage	**(-1)**	(Brooks, 1991)
Shareability	**(-1)**	(Kim, 2018)
Typicality	**(-1)**	(Brooks, 1999)
Compositionality	**(-1)**	(Brooks, 1991)
Incrementality	**(1)**	(Brooks, 1999)

(a)



(b)

Fig. 10: A focus on the desiderata concerning embodied approaches.

Figure 10 provides information on the assessment of embodied approaches. Regarding **grounding**, embodied models are certainly the artifacts that best address this desideratum as we described it. The main assumption is that cognitive tasks must always be grounded in the external environment, with concepts always connected to sensory data. This is why, as demonstrated by Rodney Brooks' automata and other seminal works (Brooks, 1991, 1999; Barsalou, 2008), perceptual-motor loops play a central role within this framework, leading us to assign a rating of **(1)** to this desideratum.

The desiderata of **coverage**, **shareability**, **typicality**, and **compositionality** are far from being key within the embodied approach. There are no projects in the literature that can be recognized as stable in these areas, even

within niche contexts. Consequently, we assigned a rating of **(-1)** for all these desiderata. Specifically, concerning coverage, embodied approaches define concepts as mere behavioural patterns, making it irrelevant to discuss concepts such as 'home', 'colour', or 'people' in this context. The coverage requirement, then, as we have described it, appears to have no particular relevance to this program. For what concerns shareability, this cannot be considered an addressable desideratum, mainly because of the anti-representational view of the embodied approaches. For what concerns typicality, within this framework, it is possible to derive behavioural patterns, but these are far from capturing prototypical or exemplar effects.[27] Regarding compositionality, the anti-representational view of the embodied approaches makes the consideration of this desideratum very puzzling (Brooks, 1991). Indeed, the conceptual representation of embodied approaches can just be inferred. Locally, concepts are not represented and there is no formalism depicting a mental model. Each activity is the random result of the composition of some processes and mechanisms.

By contrast, embodied approaches seem to be devised mainly to address **incrementality**. The grounding of cognitive tasks in perception necessitates addressing the so-called frame problem, which is the difficulty of a cognitive agent to adapt to specific and varied situations, given the nature of their internal representations (which may be abstract and rigid according to other forms of representation). An agent that is embodied in the context and environment must be capable of adaptation. In this sense, incrementality can be considered a key desideratum within the embodied framework, and the corresponding value is clearly **(1)**.

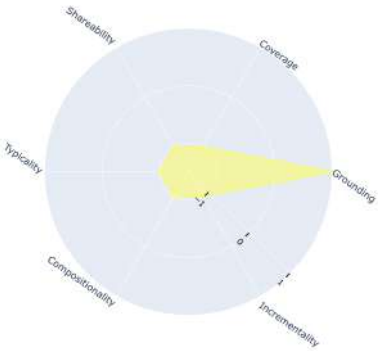
4.3 CFAI

Figure 11 and Figure 12 report the assessment values and representation for the analogical and procedural theories, respectively. **Grounding** is for sure a key priority of analogical theories research agenda. For the analogical approaches concepts are analogical representations and are taken to be mirror images of something that is outside the mind and they are considered simulations of what is perceived and experienced. In this sense, the research works in this framework always, more or less explicitly, deal with this desideratum (Roy, 2005). Similarly, the main assumption underlying proceduralism is that concepts are causally related to the external environment. However, for these latter approaches, conceptual representations are not always grounded in perception (for instance, lexical databases such as WordNet (Miller, 1995)). This motivated our choice of assigning **(1)** to analogical approaches and **(0)** to procedural approaches.

[27] Concerning this point, we are aware that E. Rosch also published a book (with Francisco Varela and Evan Thompson) entitled *The Embodied Mind* (Varela, Thompson, & Rosch, 2017), in this manuscript they address issues that could be related to the embodied program. However, Rosch's main research on basic level categories and typicality effects seems to have no established connections with this new research branch.

Desideratum	*State*	*References*
Grounding	**(1)**	(Roy, 2005)
Coverage	**(-1)**	(Gallagher & Zahavi, 2020)
Shareability	**(-1)**	(Braddon-Mitchell & Jackson, 1996), (Kim, 2018)
Typicality	**(-1)**	(Kim, 2018)
Compositionality	**(-1)**	(Kim, 2018)
Incrementality	**(-1)**	(Kosslyn et al., 2006)

(a)

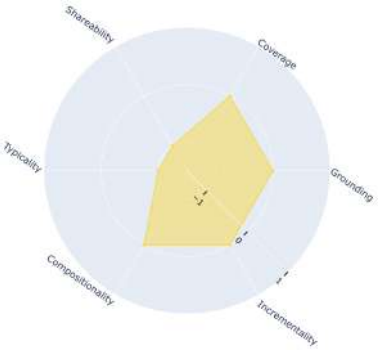


(b)

Fig. 11: A focus on the desiderata concerning analogical approaches.

Desideratum	*State*	*References*
Grounding	**(0)**	(Roy, 2005), (Fodor, 1998)
Coverage	**(0)**	(Gallagher & Zahavi, 2020), (Miller, 1995)
Shareability	**(-1)**	(Braddon-Mitchell & Jackson, 1996), (Kim, 2018)
Typicality	**(-1)**	(Fodor, 1998)
Compositionality	**(0)**	(Minsky, 1974)
Incrementality	**(0)**	(Millikan, 2004), (Fumagalli et al., 2019)

(a)



(b)

Fig. 12: A focus on the desiderata concerning procedural approaches.

Regarding **coverage**, we could not find explicit debates about this desideratum in the analogical framework literature, although some sporadic hints exist (Gallagher & Zahavi, 2020). In contrast, procedural approaches present some implementations explicitly devised for covering a wide range of concepts (see, for example, WordNet (Miller, 1995)). For this reason, we assigned a rating of **(-1)** to the former and **(0)** to the latter.

Regarding **shareability**, in the literature on analogical and procedural programs, there is evidence that concepts are not claimed to be shareable. They are internal (or local) representations/processes, and the issue of how these can be shared among agents is not explicitly addressed (Braddon-Mitchell & Jackson, 1996). For this reason, for the shareability desideratum, we assigned a rating of **(-1)** to both approaches.

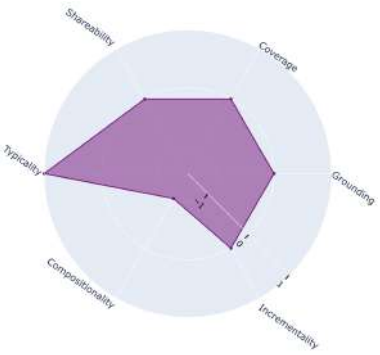
Similarly, for what concerns **typicality**, there is a lack of work in discussing this desideratum within the analogical and procedural frame. For this reason,

we assigned **(-1)** to both approaches. The same applies to the **compositionality** desideratum, which is not a requirement and is not clearly discussed in the literature of both frameworks. Still, even if this desideratum is not on the agenda of the original procedural program, it is possible to find some well-recognized computational approaches dealing with this desideratum. The work from Minsky is a prototypical example in this respect (Minsky, 1974). For this reason, we assigned **(-1)** for the analogical approaches and **(0)** for the procedural approaches.

Finally, the **incrementality** desideratum is not explicitly addressed by analogical approaches, but partially explored by procedural approaches. According to this latter paradigm, concepts can be seen as devices that change with the environment and the specific task that needs to be addressed. In this regard, see the notion of concept as *recognition ability* provided within the teleosemantics framework (Millikan, 2004; Giunchiglia & Fumagalli, 2016, 2017). For this reason, we assigned **(-1)** for the analogical approaches and **(0)** for the procedural approaches.

Desideratum	*State*	*References*
Grounding	**(0)**	(Frixione & Lieto, 2011)
Coverage	**(0)**	(Coliva, 2004; Murphy, 2004)
Shareability	**(0)**	(Frixione & Lieto, 2011)
Typicality	**(1)**	(Frixione & Lieto, 2011)
Compositionality	**(-1)**	(Frixione & Lieto, 2012)
Incrementality	**(0)**	(Frixione & Lieto, 2012; Coliva, 2004)

(a)



(b)

Fig. 13: A focus on the desiderata concerning PET approaches.

Dwelling on the desiderata covered by PET approaches (represented in Figure 13), **grounding** is often part of many debates concerning the corresponding formalisms (Frixione & Lieto, 2011). However, it would be a misrepresentation to say that this aspect is always considered whenever a corresponding representation formalism is developed. For example, in the literature related to theory-theories, the grounding issue does not seem to be the focus of attention. For this reason, we are inclined to assign a rating of **(0)**.

The range of concepts that can be represented by PET theories is wider than the one covered by analogical and procedural theories. Prototypical effects involve many common sense concepts (even abstract concepts), perhaps most of them. Indeed, even in cases where a classical definition is possible (e.g., concepts such as a *'pentagon'* or *'prime number'*) or at least a characterization in terms of necessary and sufficient conditions (e.g., *'water'*) prototypical

effects emerge. It is often this prototypical characterization that is the most relevant for many cognitive tasks (see again *'water'*) (Hampton, 2006). However, to the best of our knowledge, the current computational approaches leveraging these theories are still quite limited in covering concepts that are not related to perceptual features (e.g., key concepts such as abstract concepts and property concepts are not considered). For this reason and since there are no clear references supporting this evidence, we assigned **(-1)** for **coverage**.

PET theories account for **shareability**. Prototypical and exemplar effects, for instance, are claimed to be shareable among different agents and are often expressed by symbols that can be easily grasped by humans. However, to the best of our knowledge, this desideratum is not a priority within the PET program, and there is no widely accepted debate on how it can be properly addressed by related approaches. For this reason, we assigned a rating of **(0)** for this desideratum.

Typicality is for sure the pivotal feature here, perfectly addressed by prototypical and exemplar models. Most of the PET approaches were devised, indeed, to address this desideratum. For this reason, we assigned **(1)** in this respect.

In contrast, compositionality presents significant challenges for PET approaches. Nonetheless, various efforts have addressed compositionality within this tradition, from Hampton's work (Hampton, 2006) to conceptual spaces Gardenfors (2014) and prototype theory Lewis and Lawry (2016). Additionally, there is an ongoing debate on this topic, with studies such as Frixione and Lieto (2012) and Fodor and Lepore (2002) examining how prototypes can be compositional. Still, we have assigned a score of **(-1)**, as while these contributions exist, there is still no well-established consensus or comprehensive research program in this area.

Finally, we assigned **(0)** for the **incrementality** desideratum, which is for sure within the scope of most of the PET approaches, but its consideration seems not yet supported by a clear and shared strategy.

5 Discussion and Perspectives

The overview we provided in the previous section, suggests how the described approaches may compensate for each other in addressing the different desiderata we highlighted. For instance, from the point of view of compositionality and shareability, the GOFAI classical-symbolic approaches seem to be the ones where most efforts are concentrated. Ontologies are indeed the information artifacts that, thanks to the structure of their models, have a more immediate mapping with logical language and formal semantics. Thus, for instance, high-level cognitive tasks like planning and inferences will be best served by the applications of these models. Dually, NFAI theories, with both connectionist and embodied theories, result in being very valuable and worthwhile in addressing incrementality and grounding. Firstly, this is because their models heavily depend on "what is external" and the environment changes (i.e., inputs

coming from sensors, see previous section), where referents in the world play a key role in producing and controlling information. Secondly, this is because they rely on a characterization of concepts that is less related to the notion of "(explicit) mental representation" and more to the notions of "conceptual processing", which is essentially entangled with the mechanism of adaptation, with no exceptions. Accordingly, the PET approaches collected by the CFAI theories can provide an exceptional account of what we have described as "typicality". This essential desideratum can be perfectly addressed, for instance, by the prototypical and the exemplar theories.

Consistently, these different representations of concepts should not be taken as competing with each other. Taking inspiration from what some authors are already claiming (see for instance Ruth G. Millikan (Millikan, 2017)), we may say that each conceptual representation has an *ad hoc* role in supporting a specific function within an (artificial) cognitive system, and, possibly, referring to a specific category of objects. In line with this, what makes it possible to identify a computational representation of a concept, as well as a certain structure and certain properties, is its function within an artificial cognitive system and, more generally, its function in explaining behaviour that is considered "intelligent". For instance, a concept about things like books or apples can be something in the system used to deal with those specific things according to a certain goal, e.g., *to recognize* books (from images or language) or *to reason* about apples (as fruits in general or as food). For this reason, we can also say that concepts are representations that enable abilities (Giunchiglia & Fumagalli, 2016; Fumagalli, Ferrario, & Guizzardi, 2024), and all the desiderata we reviewed so far can be seen as opportunities to enable those abilities and address specific tasks.

A discussion like the one addressed in this work could serve to enhance our understanding of the most suitable representation for a given task and referent. For example, in (Millikan, 2000), Millikan asserts that concepts categorized as "substance", potentially resembling the structure of neural networks (Fumagalli et al., 2019; Millikan, 2017), work as essential tools for recognizing physical objects. The computational challenge lies in devising a system capable of accommodating diverse representation types, thereby encompassing various desiderata and corresponding abilities. The distinctions between GOFAI, NFAI, and CFAI, as highlighted earlier, offer several insights. Fundamental questions persist. *i)* Which computational representation of concepts is optimal for supporting specific artificial cognitive tasks (e.g., learning)? *ii)* How should we consider the computational representation of a concept w.r.t. other existing representations that support different yet crucial and complementary artificial cognitive tasks (e.g., reasoning)?

In this sense, the analysis presented in this paper could also support a cognitively-inspired design of computational approaches. Moreover, it can be used to support the comparison of existing technologies, considering them according to their potential (artificial-)cognitive purpose and their relevance concerning the discussed desiderata. For example:

(1) Knowing that an approach falls into a specific category can lead to greater awareness of the desiderata on which it may be most reliable. For instance, knowing that neural networks fall under the NFAI program allows us to recognize their strength in incrementality rather than compositionally;
(2) Understanding which desiderata the approaches are most reliable for can facilitate more timely validation of these approaches. Keeping in mind the different cognitive inspirations of ontologies and neural networks makes it easier to identify tasks, such as reasoning processes or learning capabilities, where direct comparison might be less meaningful;
(3) While designing an artificial intelligence (hybrid) system, it might become easier to discern which technology to adopt for each cognitive task. For instance, when creating a system that can plan movements in space and gather information from the environment, it would be reasonable to leverage the features of GOFAI and embodied approaches.

Concerning item *(3)*, it is important to point out that in the artificial intelligence scenario, there are already a lot of attempts to embed and integrate representations of concepts. In recent years, for example, there has been a lot of talk about integrating symbolic and sub-symbolic AI, where the main effort is to combine knowledge graphs such as ontologies with machine learning models. In (Serafini & d'Avila Garcez, 2016; Fumagalli & Giunchiglia, 2020; Monka, Halilaj, & Rettinger, 2022), for example, data structures encoding logical theories are adopted for improving the performance of the learning capabilities of some given models, to improve their cross-domain (i.e., generalization) capabilities and allow to decrease the amount of data needed to address a prediction task in a reasonably accurate manner. Machine learning and connectionist approaches are also used to enable the induction of some structural knowledge that can be used then for reasoning tasks (Nickel, Murphy, Tresp, & Gabrilovich, 2015; Fumagalli, Sales, & Guizzardi, 2021; Fumagalli, Sales, Baião, & Guizzardi, 2022). A huge debate is going on about integrating knowledge graphs with large language models (Yasunaga, Ren, Bosselut, Liang, & Leskovec, 2021; Pan et al., 2024), to improve the reasoning capabilities of the latter and overcome the rigidity of the former.

However, these attempts, most of the time, are purely engineering solutions that do not consult the literature devoted to concept analysis, which, as we have discussed, presents several pivotal insights to defining the requirements needed to support so-called "intelligent behaviours". In this respect, an awareness of what the efforts in philosophy and psychology have been in delineating the characteristics of concepts can be a valuable support, mainly in two respects. Firstly, we would say, as a normative support, since this can enable a more critical attitude in using certain representations for tasks that are not the most appropriate. For example, lately, there has been an effort in AI to understand how connectionist models can be used in reasoning tasks (Webb, Holyoak, & Lu, 2023). To what extent does this effort make sense, given that CFAI models, and not NFAI models, were designed for exactly this? The literature about concepts could be then useful in understanding a kind

of *proper functioning* of the adopted formalism. Secondly, we would say, as an integration support. In the conceptual analysis, there have been attempts to combine different approaches, clearly motivated by the tasks to be addressed. For instance, part of the CFAI theories (see for instance some hybrid procedural approaches like the one in Peng, Lu, Li, and Wong (2015)) provide interesting insights on how to merge some of the features of the GOFAI and NFAI approaches. The works described in (Shavlik, 1994), (Sun & Alexandre, 2013), are some rather successful experiments in this direction. These attempts could be replicated somehow in a computational setting, to efficiently cover the typicality, grounding, and compositionality desideratum in a hybrid solution.

Moreover, as mentioned in the introduction, this work can also contribute to philosophy, particularly within the field of conceptual engineering, including efforts to re-engineer concepts for ameliorative purposes. To provide more context, the process of conceptual re-engineering typically involves four components: *description*, *evaluation*, *improvement*, and *implementation* (Isaac et al., 2022). Understanding which paradigm of conceptual representation is being employed in a given case can help refine the goals, target, and methodology. For example, consider Haslanger's (Haslanger, 2000) project to revise the concept of 'woman' in order to combat social injustice. Haslanger argues that the concept of a woman includes being a person who is systematically subordinated based on perceived or imagined female bodily characteristics. Haslanger's claim may reflect a typicality effect that one aims to modify. In such a case, recognizing that this is a PET paradigm of concept representation (the target) would guide the design of the goal and the strategy to achieve it—e.g., addressing not the necessary or sufficient conditions, but rather the undesired typicality effect.

Finally, it is crucial to highlight that the reciprocal benefits are noteworthy. Artificial Intelligence (AI) research not only reaps advantages from conceptual analysis but also plays a pivotal role in testing and potentially generating new theories within this domain. The unprecedented convergence of purposes across disciplines — ranging from AI to philosophy of mind, psychology and neuroscience — is evident in this exploration of conceptual representations. This journey through conceptual representations can serve as another concrete example of the intersection between these fields. The aspiration is that a continuous synergy between these disciplines will not only propel advancements in each but also deepen our comprehension of what constitutes "intelligent behaviour".

6 Conclusion

In this paper, we presented a classification and description of the major existing computational approaches to the representation of concepts. In the proposed review we classified the major works according to the reference background theories, i.e., GOFAI, NFAI, and CFAI theories. Moreover, we

set the stage for a more cognitively-inspired discussion on the computational representation of concepts.

The main purpose of this effort was to provide a reader who may not be familiar with theories of concepts with an introduction to major themes in this research and with pointers to different research projects. Furthermore, the contribution can be taken as a support for bridging different disciplinary fields, such as philosophy of mind, psychology, and artificial intelligence and, also, as a first step towards an assessment of some of the most remarkable representational approaches. Finally, another intention of ours was to trigger combined future research in AI, philosophy of mind and psychology. Accordingly, we look forward to encouraging the exploitation and evolution of the categorization and assessment we proposed, paving the way to new possible integrated theories and advancements in the research on concepts.

References

Adams, F., & Aizawa, K. (2010). Causal theories of mental content.

Baader, F., Calvanese, D., McGuinness, D., Patel-Schneider, P., & Nardi, D. (2003). *The description logic handbook: Theory, implementation and applications*.

Baader, F., & Gil, O. F. (2024). Extending the description logic el with threshold concepts induced by concept measures. *Artificial Intelligence*, *326*, 104034.

Baader, F., Horrocks, I., Lutz, C., & Sattler, U. (2017). *Introduction to description logic*.

Bandrowski, A., Brinkman, R., Brochhausen, M., Brush, M. H., Bug, B., Chibucos, M. C., ... others (2016). The ontology for biomedical investigations. *PloS one*, *11*(4), e0154556.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and brain sciences*, *22*(4), 577–660.

Barsalou, L. W. (2008). Grounded cognition. *Annu. Rev. Psychol.*, *59*(1), 617–645.

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

Bengio, Y., Lamblin, P., Popovici, D., & Larochelle, H. (2006). Greedy layer-wise training of deep networks. *Advances in neural information processing systems*, *19*.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. *Scientific american*, *284*(5), 34–43.

Berthold, M. R., Sudweeks, F., Newton, S., & Coyne, R. D. (1998). It makes sense: Using an autoassociative neural network to explore typicality in computer mediated discussions.

Besnard, P. (2013). *An introduction to default logic*.

Borgo, S., Ferrario, R., Gangemi, A., Guarino, N., Masolo, C., Porello, D., ... Vieu, L. (2022, January). DOLCE: A descriptive ontology for linguistic

and cognitive engineering. *Applied Ontology*, *17*(1), 45–69. doi: 10.3233/AO-210259

Bottazzi, E., Ferrario, R., & Masolo, C. (2012). The mysterious appearance of objects. In *Formal ontology in information systems* (pp. 59–72).

Braddon-Mitchell, D., & Jackson, F. (1996). Philosophy of mind and cognition: An introduction.

Brooks, R. A. (1991). Intelligence without representation. *Artificial intelligence*, *47*(1-3), 139–159.

Brooks, R. A. (1999). *Cambrian intelligence: The early history of the new ai.*

Broomhead, D. S., & Lowe, D. (1988). *Radial basis functions, multi-variable functional interpolation and adaptive networks* (Tech. Rep.). Royal Signals and Radar Establishment Malvern (United Kingdom).

Byrne, R. M. (2023). Good explanations in explainable artificial intelligence (xai): Evidence from human explanatory reasoning. In *Ijcai* (pp. 6536–6544).

Cangelosi, A., Coventry, K. R., Rajapakse, R., Joyce, D., Bacon, A., Richards, L., & Newstead, S. N. (2005). Grounding language in perception: A connectionist model of spatial terms and vague quantifiers. In *Modeling language, cognition and action* (pp. 47–56).

Cangelosi, A., Greco, A., & Harnad, S. (2000). From robotic toil to symbolic theft: grounding transfer from entry-level to higher-level categories1. *Connection science*, *12*(2), 143–162.

Chalmers, D. J. (2020). What is conceptual engineering and what should it be? *Inquiry*, 1–18.

Chen, X., Jia, S., & Xiang, Y. (2020). A review: Knowledge reasoning over knowledge graph. *Expert Systems with Applications*, *141*, 112948.

Coliva, A. (2004). I concetti: teorie ed esercizi.

Craver, C. F., & Bechtel, W. (2007). Top-down causation without top-down causes. *Biology & philosophy*, *22*, 547–563.

Davis, E., & Marcus, G. (2015). Commonsense reasoning and commonsense knowledge in artificial intelligence. *Communications of the ACM*, *58*(9), 92–103.

Davis, T., & Poldrack, R. A. (2014). Quantifying the internal structure of categories using a neural typicality measure. *Cerebral Cortex*, *24*(7), 1720–1737.

Fine, T. L. (1999). *Algorithms for designing feedforward networks.*

Fodor, J. A. (1998). *Concepts: Where cognitive science went wrong.*

Fodor, J. A. (2008). *Lot 2: The language of thought revisited.*

Fodor, J. A., & Lepore, E. (2002). *The compositionality papers.*

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*(1-2), 3–71.

Frixione, M., & Lieto, A. (2011). Representing concepts in artificial systems: a clash of requirements. *Proc. HCP*, 75–82.

Frixione, M., & Lieto, A. (2012). Representing concepts in formal ontologies. compositionality vs. typicality effects. *Logic and Logical Philosophy*,

21(4), 391–414.

Fumagalli, M., Bella, G., & Giunchiglia, F. (2019). Towards understanding classification and identification. In *Pricai 2019: Trends in artificial intelligence: 16th pacific rim international conference on artificial intelligence, cuvu, yanuca island, fiji, august 26–30, 2019, proceedings, part i 16* (pp. 71–84).

Fumagalli, M., Ferrario, R., & Guizzardi, G. (2024). A teleological approach to information systems design. *Minds and Machines*, *34*(3), 23.

Fumagalli, M., & Giunchiglia, F. (2020). Ontology-driven cross-domain transfer learning. In *Formal ontology in information systems: Proceedings of the 11th international conference (fois 2020)* (Vol. 330, p. 249).

Fumagalli, M., Sales, T. P., Baião, F. A., & Guizzardi, G. (2022). Conceptual model visual simulation and the inductive learning of missing domain constraints. *Data & Knowledge Engineering*, *140*, 102040.

Fumagalli, M., Sales, T. P., & Guizzardi, G. (2021). Mind the gap!: Learning missing constraints from annotated conceptual model simulations. In *The practice of enterprise modeling: 14th ifip wg 8.1 working conference, poem 2021, riga, latvia, november 24–26, 2021, proceedings 14* (pp. 64–79).

Gallagher, S., & Zahavi, D. (2020). *The phenomenological mind.*

Gardenfors, P. (2014). *The geometry of meaning: Semantics based on conceptual spaces.*

Genesereth, M. R., & Nilsson, N. J. (2012). *Logical foundations of artificial intelligence.* Morgan Kaufmann.

Giunchiglia, F., & Fumagalli, M. (2016). Concepts as (recognition) abilities. In *Fois* (pp. 153–166).

Giunchiglia, F., & Fumagalli, M. (2017). Teleologies: Objects, actions and functions. In *Conceptual modeling: 36th international conference, er 2017, valencia, spain, november 6–9, 2017, proceedings 36* (pp. 520–534).

Gruber, T. (1993). A translation approach to portale ontologies knowledge acquisition. *Disponıvel em: http://ksl-web. stanford. edu/KSL Abstracts/KSL-92-71. html. Acesso em*, *10*(09), 2004.

Guarino, N., Oberle, D., & Staab, S. (2009). What is an ontology? *Handbook on ontologies*, 1–17.

Guarino, N., & Welty, C. (2002). Evaluating ontological decisions with ontoclean. *Communications of the ACM*, *45*(2), 61–65.

Guha, R. V., Brickley, D., & Macbeth, S. (2016). Schema. org: evolution of structured data on the web. *Communications of the ACM*, *59*(2), 44–51.

Guizzardi, G., Botti Benevides, A., Fonseca, C. M., Porello, D., Almeida, J. P. A., & Prince Sales, T. (2022). Ufo: Unified foundational ontology. *Applied ontology*, *17*(1), 167–210.

Guizzardi, G., & Guarino, N. (2024). Explanation, semantics, and ontology. *Data & Knowledge Engineering*, 102325.

Hampton, J. A. (2006). Concepts as prototypes. *Psychology of learning and*

motivation, *46*, 79–113.

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, *42*(1-3), 335–346.

Haslanger, S. (2000). Gender and race:(what) are they?(what) do we want them to be? *Noûs*, *34*(1), 31–55.

Haugeland, J. (1989). *Artificial intelligence: The very idea*.

Hayes, B., et al. (2013). First links in the markov chain. *American Scientist*, *101*(2), 252.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 770–778).

Hebb, D. O. (2005). *The organization of behavior: A neuropsychological theory*.

Hofweber, T. (2024). Inescapable concepts. *Australasian Journal of Philosophy*, *102*(1), 159–179.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, *79*(8), 2554–2558.

Hupkes, D., Dankers, V., Mul, M., & Bruni, E. (2020). Compositionality decomposed: How do neural networks generalise? *Journal of Artificial Intelligence Research*, *67*, 757–795.

Isaac, M. G., Koch, S., & Nefdt, R. (2022). Conceptual engineering: A road map to practice. *Philosophy Compass*, *17*(10), e12879.

Jackson, D. (2021). *The essence of software: Why concepts matter for great design*.

Janowicz, K., & Compton, M. (2010). The stimulus-sensor-observation ontology design pattern and its integration into the semantic sensor network ontology. In *Ssn*.

Johnson-Laird, P. N. (1977). Procedural semantics. *Cognition*, *5*(3), 189–214.

Joyce, D., Richards, L., Cangelosi, A., & Coventry, K. R. (2003). On the foundations of perceptual symbol systems: Specifying embodied representations via connectionism. In *The logic of cognitive systems: Proceedings of the fifth international conference on cognitive modeling* (pp. 147–152).

Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of artificial intelligence research*, *4*, 237–285.

Kansky, K., Silver, T., Mély, D. A., Eldawy, M., Lázaro-Gredilla, M., Lou, X., ... George, D. (2017). Schema networks: Zero-shot transfer with a generative causal model of intuitive physics. In *International conference on machine learning* (pp. 1809–1818).

Kim, J. (2018). *Philosophy of mind*. Routledge.

Klir, G., & Yuan, B. (1995). *Fuzzy sets and fuzzy logic* (Vol. 4).

Kosslyn, S. M., Thompson, W. L., & Ganis, G. (2006). *The case for mental imagery*. Oxford University Press.

Kuhn, W. (2009). A functional ontology of observation and measurement. In

Geospatial semantics: Third international conference, geos 2009, mexico city, mexico, december 3-4, 2009. proceedings 3 (pp. 26–43).

Laurence, S., & Margolis, E. (1999). Concepts and cognitive science.

Lenat, D. (2022). Creating a 30-million-rule system: Mcc and cycorp. *IEEE Annals of the History of Computing*, *44*(1), 44–56.

Levesque, H. J. (1986). Knowledge representation and reasoning. *Annual review of computer science*, *1*(1), 255–287.

Lewis, M., & Lawry, J. (2016). Hierarchical conceptual spaces for concept combination. *Artificial Intelligence*, *237*, 204–227.

Lieto, A., Minieri, A., Piana, A., & Radicioni, D. P. (2015). A knowledge-based system for prototypical reasoning. *Connection Science*, *27*(2), 137–152.

Lieto, A., & Pozzato, G. L. (2018). A description logic of typicality for conceptual combination. In *International symposium on methodologies for intelligent systems* (pp. 189–199).

Lieto, A., Radicioni, D. P., & Rho, V. (2015). A common-sense conceptual categorization system integrating heterogeneous proxytypes and the dual process of reasoning. In *Twenty-fourth international joint conference on artificial intelligence.*

Margolis, E., & Laurence, S. (2023). Concepts. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy* (Fall 2023 ed.). https://plato.stanford.edu/archives/fall2023/entries/concepts/.

Margolis, E. E., & Laurence, S. E. (1999). *Concepts: core readings.*

Masolo, C., Botti Benevides, A., & Porello, D. (2018). The interplay between models and observations. *Applied Ontology*, *13*(1), 41–71.

McClelland, J. L., & Rogers, T. T. (2003). The parallel distributed processing approach to semantic cognition. *Nature reviews neuroscience*, *4*(4), 310–322.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, *5*, 115–133.

Medsker, L. R., & Jain, L. (2001). Recurrent neural networks. *Design and Applications*, *5*(64-67), 2.

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, *38*(11), 39–41.

Millikan, R. G. (2000). *On clear and confused ideas: An essay about substance concepts.* Cambridge University Press.

Millikan, R. G. (2004). *Varieties of meaning: the 2002 jean nicod lectures.*

Millikan, R. G. (2017). *Beyond concepts: Unicepts, language, and natural information.*

Minsky, M. (1974). *A framework for representing knowledge.* MIT, Cambridge.

Monka, S., Halilaj, L., & Rettinger, A. (2022). A survey on visual transfer learning using knowledge graphs. *Semantic Web*, *13*(3), 477–510.

Murphy, G. (2004). *The big book of concepts.*

Nickel, M., Murphy, K., Tresp, V., & Gabrilovich, E. (2015). A review of relational machine learning for knowledge graphs. *Proceedings of the*

IEEE, *104*(1), 11–33.

Noy, N. F., & Klein, M. (2004). Ontology evolution: Not the same as schema evolution. *Knowledge and information systems*, *6*, 428–440.

Olsen, K. (2020). *Neuroevolution of artificial general intelligence* (Unpublished master's thesis).

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

Ortmann, J., & Daniel, D. (2011). An ontology design pattern for referential qualities. In *International semantic web conference* (pp. 537–552).

Otte, J. N., Beverley, J., & Ruttenberg, A. (2022). Bfo: Basic formal ontology. *Applied ontology*, *17*(1), 17–43.

Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J., & Wu, X. (2024). Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*.

Papavasileiou, E., Cornelis, J., & Jansen, B. (2021). A systematic literature review of the successors of "neuroevolution of augmenting topologies". *Evolutionary computation*, *29*(1), 1–73.

Peng, B., Lu, Z., Li, H., & Wong, K.-F. (2015). Towards neural network-based reasoning. *arXiv preprint arXiv:1508.05508*.

Pezzulo, G., Barsalou, L. W., Cangelosi, A., Fischer, M. H., McRae, K., & Spivey, M. J. (2013). Computational grounded cognition: a new alliance between grounded cognition and computational modeling. *Frontiers in psychology*, *3*, 612.

Piantadosi, S. T., Muller, D. C., Rule, J. S., Kaushik, K., Gorenstein, M., Leib, E. R., & Sanford, E. (2024). Why concepts are (probably) vectors. *Trends in Cognitive Sciences*.

Porello, D., Kutz, O., Righetti, G., Troquard, N., Galliani, P., & Masolo, C. (2019). A toothful of concepts: Towards a theory of weighted concept combination. In M. Simkus & G. E. Weddell (Eds.), *Proceedings of the 32nd international workshop on description logics, oslo, norway, june 18-21, 2019* (Vol. 2373). CEUR-WS.org. Retrieved from https://ceur-ws.org/Vol-2373/paper-24.pdf

Prinz, J. J. (2004). *Furnishing the mind: Concepts and their perceptual basis*.

Probst, F. (2008). Observations, measurements and semantic reference spaces. *Applied Ontology*, *3*(1-2), 63–89.

Quillian, M. R. (1967). Word concepts: A theory and simulation of some basic semantic capabilities. *Behavioral science*, *12*(5), 410–430.

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

Rosch, E. (1978). Principles of categorization. In *Cognition and categorization* (pp. 27–48).

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, *65*(6), 386.

Roy, D. (2005). Grounding words in perception and action: computational

insights. *Trends in cognitive sciences*, *9*(8), 389–396.

Russell, S. J., & Norvig, P. (2010). Artificial intelligence a modern approach.

Santoro, A., Raposo, D., Barrett, D. G., Malinowski, M., Pascanu, R., Battaglia, P., & Lillicrap, T. (2017). A simple neural network module for relational reasoning. *Advances in neural information processing systems*, *30*.

Serafini, L., & d'Avila Garcez, A. S. (2016). Learning and reasoning with logic tensor networks. In *Conference of the italian association for artificial intelligence* (pp. 334–348).

Shavlik, J. W. (1994). Combining symbolic and neural learning. *Machine Learning*, *14*, 321–331.

Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, *171*(3972), 701–703.

Siegel, N., Goolsbey, K., Kahlert, R., & Matthews, G. (2004). The cyc system: Notes on architecture. *Cycorp, Inc*, *9*.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., ... others (2016). Mastering the game of go with deep neural networks and tree search. *nature*, *529*(7587), 484–489.

Sloman, A. (2014). How can we reduce the gulf between artificial and natural intelligence? In *Aic* (pp. 1–13).

Smith, B. (2001). Beyond concepts: ontology as reality representation.

Smith, E. E., Osherson, D. N., Rips, L. J., & Keane, M. (1988). Combining prototypes: A selective modification model. *Cognitive science*, *12*(4), 485–527.

Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and brain sciences*, *11*(1), 1–23.

Studer, R., Benjamins, V. R., & Fensel, D. (1998). Knowledge engineering: Principles and methods. *Data & knowledge engineering*, *25*(1-2), 161–197.

Sun, R., & Alexandre, F. (2013). *Connectionist-symbolic integration: From unified to hybrid approaches.*

Tyler, L. K., Moss, H. E., Durrant-Peatfield, M., & Levy, J. (2000). Conceptual structure and the structure of concepts: A distributed account of category-specific deficits. *Brain and language*, *75*(2), 195–231.

Vandenbussche, P.-Y., Atemezing, G. A., Poveda-Villalón, M., & Vatant, B. (2017). Linked open vocabularies (lov): a gateway to reusable semantic vocabularies on the web. *Semantic Web*, *8*(3), 437–452.

Varela, F. J., Thompson, E., & Rosch, E. (2017). The embodied mind. *(No Title)*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Vernon, D., Metta, G., & Sandini, G. (2007). 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.

Way, E. C. (1997). Connectionism and conceptual structure. *American Behavioral Scientist*, *40*(6), 729–753.

Webb, T., Holyoak, K. J., & Lu, H. (2023). Emergent analogical reasoning in large language models. *Nature Human Behaviour*, *7*(9), 1526–1541.

Wittgenstein, L. (2009). *Philosophical investigations*.

Yasunaga, M., Ren, H., Bosselut, A., Liang, P., & Leskovec, J. (2021). Qa-gnn: Reasoning with language models and knowledge graphs for question answering. *arXiv preprint arXiv:2104.06378*.

Yeung, C.-m. A., & Leung, H.-f. (2006). Formalizing typicality of objects and context-sensitivity in ontologies. In *Proceedings of the fifth international joint conference on autonomous agents and multiagent systems* (pp. 946–948).