Minds and Machines (2023) 33:397–427
https://doi.org/10.1007/s11023-023-09638-w

An Alternative to Cognitivism: Computational Phenomenology for Deep Learning

Pierre Beckmann[1] · **Guillaume Köstner**[1] · **Inês Hipólito**[2,3]

Received: 15 February 2023 / Accepted: 2 June 2023 / Published online: 29 June 2023

Abstract

We propose a non-representationalist framework for deep learning relying on a novel method computational phenomenology, a dialogue between the first-person perspective (relying on phenomenology) and the mechanisms of computational models. We thereby propose an alternative to the modern cognitivist interpretation of deep learning, according to which artificial neural networks encode representations of external entities. This interpretation mainly relies on neuro-representationalism, a position that combines a strong ontological commitment towards scientific theoretical entities and the idea that the brain operates on symbolic representations of these entities. We proceed as follows: after offering a review of cognitivism and neuro-representationalism in the field of deep learning, we first elaborate a phenomenological critique of these positions; we then sketch out computational phenomenology and distinguish it from existing alternatives; finally we apply this new method to deep learning models trained on specific tasks, in order to formulate a conceptual framework of deep-learning, that allows one to think of artificial neural networks' mechanisms in terms of lived experience.

Keywords Computational sciences · Deep learning · Phenomenology · Cognitivism · Cognitive science · Neuroscience

✉ Pierre Beckmann
pierre.beckmann@unil.ch

[1] University of Lausanne, Lausanne, Switzerland

[2] Berlin School of Mind and Brain, Humboldt-Universität zu Berlin, Berlin, Germany

[3] Department of Philosophy, Macquarie University, Sydney, NSW, Australia

1 Introduction

In the past years, deep learning (DL) has achieved impressive feats, its artificial neural networks (ANNs) competing with human performance on tasks involving the understanding or the generation of text (Devlin, et al., 2019; Brown, et al., 2020; Schulman, et al., 2022), images (Radford, et al., 2021; Ramesh, et al., 2021) or speech (Baevski, et al., 2020; Hsu, et al., 2021). DL has had a revolutionary impact in the industry and society, as well as in scientific research (DeVries, et al., 2018; Davies, et al., 2021; Jumper, et al., 2021). Specifically, ANNs' ability to successfully mimic a number of human cognitive processes, has fueled comparative research between DL and cognitive sciences or neurosciences (Yamins, 2016; Kumar, et al., 2022; McClelland, 2022; Millet, et al., 2022). Recently, DL also motivated renewed philosophical perspectives upon old questions relating to the mind, brain and behavior (Buckner, 2019; Sloman, 2019; Fazi, 2021; Perconti & Plebe, 2020).

Despite its successes, ANNs are notoriously hard to interpret, in the sense that we cannot exactly understand how they solve their tasks (Boge, 2022). For this reason they are sometimes referred to as "black boxes" (Castelvecchi, 2016). This opacity makes DL models susceptible to diverse interpretations through different conceptual frameworks. The most prominent framework for the interpretation of DL has been cognitivism, the first research program in cognitive sciences (MacKay, et al., 1956; Lees & Chomsky, 1957; Minsky, 1961). Relying on the functioning of the Turing machine, cognitivism defends cognition in terms of *symbol* manipulation: cognitive processes are thought to rely on *representation* of entities[1] of an external pregiven world. This approach has been influential both in cognitive sciences and philosophy of mind, with Fodor's computational theory of mind being the most prominent (Fodor, 1983).

DL originates within connectionism, a computationalist framework that disputes that cognitive, computational processes are leveraged by symbol manipulation (McCulloch & Pitts, 1943; Rosenblatt, 1958; Rumelhart, et al., 1986). Aiming for a more "biological" resemblance of the distributed operations of the brain, connectionism brings forth the ancestors of DL models, such as Rosenblatt's perceptron (Rosenblatt, 1958). Unlike cognitivism's Turing machine, these mathematical models do not need to be implemented fully: they can learn how to solve tasks by slowly adjusting to new inputs. With these models, connectionism promotes a new conceptual framework to think about human cognition as *emergence* of global *states* to fulfill cognitive functions. It therefore initially opposes cognitivist's symbol-like representations : "one-to-one mappings" between entity and representation (Rosenblatt, 1958)[2].

[1] We use the ontologically neutral term of "entity" as it does not matter for our purpose whether the ontology of the world is conceived in terms of objects, properties, relations, processes, events, tropes, or other metaphysical categories. What is important is the representative relation holding between those entities and the mental symbols according to cognitivism. It should also be noted that this relation does not imply that the grammar of those symbols has to match the ontology of the world. For example, the representation *of* a process does not need to represent it *as a process*.

[2] However, as Bechtel and Abrahamsen (1991) notes, from the 1980s, connectionists try to reconcile their theory with the predominant cognitivism, insisting that their models "should be embraced as a subsymbolic alternative to symbolic models of cognition".

Although it stems from connectionism, DL is today largely thought in terms of *representation*-based operations – relying on the cognitivist toolkit. In fact, connectionist models' self-organizational capabilities happen to be of use for the convinced cognitivist, because they can provide the otherwise unanswered explanation of how a system can *learn* symbols. ANNs learn symbolic representations of external properties on the basis of which they can execute further computations to solve tasks. This stance is quickly associated with neuro-representationalism (NR), combining a strong realism towards scientific entities with the idea that we experience a brain generated model of these entities (Churchland & Sejnowski, 1990; Milkowski, 2013; Mrowca, et al., 2018; Sitzmann, et al., 2020).

NR with the notion of representation is pervasive both in computational neuroscience (Piantadosi, 2021; Poldrack, 2021) and DL itself (LeCun, et al., 2015; Ha & Schmidhuber, 2018; Gidaris, et al., 2018; Chen, et al., 2020; Goh, et al., 2021; Matsuo, et al., 2022). In *Nature*'s most cited paper, "Deep Learning" for example, ANNs are described from the first paragraph, as machines that can "be fed with raw data" and "automatically discover the *representations* needed for detection or classification" (LeCun, et al., 2015). To the best of our knowledge, this framework is implicitly prevalent in deep learning literature. Often, deep learning researchers motivate their use of the cognitivist concept of representation by relying on NR; see this recent deep learning textbook for example:

> More often than not, hidden layers have fewer neurons than the input layer to force the network to learn compressed representations of the original input. For example, while our eyes obtain raw pixel values from our surroundings, our brain thinks in terms of edges and contours. This is because the hidden layers of biological neurons in our brain force us to come up with better representations for everything we perceive. (Buduma et al., 2022)

Furthermore, DL researchers commonly interpret ANNs as learning "world models", that mimic external world structures and dynamics to plan ahead (Ha & Schmidhuber, 2018; Matsuo, et al., 2022). When used to understand the mind, these "world models" are oftentimes reduced to perception; the idea being the following: because perception is indirect, the brain must build internal models in an attempt to represent what could potentially be the perceptual space state of affairs (Von der Malsburg, 1995; Ashby, 2014; Saddler, et al., 2021). This introduction of intermediate representations posits the existence of an external reality, the ontological structure of which can (arguably) be known independently of the way the mind relates to it (metaphysical realism), but that we can never directly access through our perception (representationalism). In philosophy, this form of representationalism is famously opposed by phenomenology, which puts on hold the question of the existence of an external reality in favor of a rigorous description of lived experience (Husserl, 1931; Merleau-Ponty, 1945). Today, in cognitive science, and under the influence of phenomenology, representationalism is further challenged in the Embodied and Enactive Cognitive Science (EECS) research program (Varela, et al., 1991; Hutto & Myin, 2012; Chemero, 2011; Di Paolo, et al., 2017; Gallagher, 2017).

Taking into consideration these critiques of representationalism, this paper aims to provide a conceptual framework of DL that does not rely on symbols as the basic units of cognition. As such, we chose to rely on phenomenology, privileging insights from the careful description of reality as it appears to us in first-person perspective. Furthermore, following Lutz and Thompson (2003), we will leverage three levels of enquiry, or sources of exploration of cognition : *the neurophysiological source*, *the phenomenological source* and *the computational source*. Historically, the computational source was used to formalize and link the findings obtained from the two other sources. With DL successfully mimicking some of our cognitive processes, the computational source now generates new *data* and becomes a new source of exploration. This observation opens up the possibility of *computational phenomenology (CP)*: an exclusive dialogue between the *computational* and the *phenomenological* source that puts on hold the question of the material basis of cognitive processes (which belongs to the neurophysiological source). The point of such a dialogue is not to disqualify the neurophysiological source, but rather to provisionally let the two other sources free of any constraint or import coming from the third one. As such, this dialogue is more faithful to phenomenology (relying on first-person descriptions of experience) than EECS approaches that tend to recast phenomenology in a more naturalist, third-person point of view. Turning to DL, we find that from lived experience, the apparent non-decomposability of ANN operations – their "black box" aspect – is not surprising as the underlying mechanics of many of our cognitive processes are unclear, or opaque, to us. This observation allows us to both propose new phenomenology-drawn concepts to think of ANN operations, and to embrace the opaque processual nature of cognitive processes.

It should be noted that the proposition contained in the following pages does not intend to constitute a research program but rather the outlines of such a program. As such, we are not suggesting that CP is the only valid approach to ANNs that ought to be adopted by researchers. At the preliminary stage we find ourselves in the development of CP, it can only claim the status of an alternative conceptual framework to interpretations already present in the field of research. Thus, the contribution of this paper is threefold, we propose computational phenomenology, a new methodology and conceptual framework for philosophers and cognitive scientists to conceive (1) of the mind and its relation to (2) task-solving computational models; such a conceptual framework provides (3) DL engineers with a new experience-based toolkit by applying this methodology to the operations of ANNs.

2 Cognitivism and Neuro-representationalism in Deep Learning

DL is concerned with the design and the training of ANNs. An ANN sequentially connects *layers* of (non-linear) threshold-activated nodes with linear operations according to a set of *weights* to transform an input into an output. Layers between the input layer and the output layer are called *hidden layers*; the "deep" in DL refers to the multiple hidden layers in the ANNs.

The weights of an ANN are not fixed but are gradually adjusted to better solve a precise task. To enable learning, the deep learning researcher picks three main

ingredients, given some particular data: the network's *architecture* (the way in which the different nodes are connected through the weights), a *loss function* (that models the objective of the task) and a *learning rule* (that dictates the way the weights are updated according to the loss function – and is almost always based on an algorithm called *backpropagation* (Rumelhart, et al., 1986). Once the optimization of the weights – also called *training* – is done, ANNs allow *inference*: the propagation of new input all the way to output layers, causing the nodes of the hidden layers to *activate* in a particular way. In short, once the ANN is trained, the *weights* are fixed and allow the processing of new inputs; on the other hand, the *activations* are obtained through inference and always correspond to one given input. These weight-determined successive activations, or patterns of activations, are not easily interpretable. Importantly, this means that it is not possible – so far at least – to clearly decompose them into explainable steps, or symbol-based operations. The original connectionist framework sees them as nothing more than *emerging states* and is already satisfied that they allow solving a particular cognitive task (Varela, et al., 1991, p. 98). However, some researchers reject the apparent weaker explanatory power of this interpretation in favor of reading of DL in terms of cognitivist's mind-computer metaphor : i.e. conceiving cognition as symbolic computation (LeCun, et al., 2015; Buduma, et al., 2022; Matsuo, et al., 2022).

A cognitivist reading of an ANN relies on interpreting its pattern of activations as symbol-based operations. We argue that this reading isn't motivated by technical reasons but that it is grounded in a philosophical worldview. Cognitivism, from its birth in the 1950s, relies on the mechanisms of the Turing machine to interpret human cognition (Putnam, 1967; Fodor, 1983). When one implements a Turing machine, one decides what a one or a zero in a given cell *means*, that is to say, to what it relates to in our world. This physically manipulable entry then carries a human-assigned meaning; it becomes a *symbol* that *represents* a given external entity, and on the basis of which Turing machines can carry out meaningful operations. In a second step, cognitivism transposes this functioning to the mind. The mind is thought to run on the basis of *mental representations*, that semantically *encode* properties of an external pregiven world analogously to the symbols of a Turing machine[3]. Cognitivism therefore implies a *metaphysical realism*, the view that there exists an external world, with entities, independently of our perception or thoughts of it. Cognitivism can be systematized around the answers given to three major questions:

1) What is The World?
   An external reality that exists independently of our cognition of it.
2) What is A Representation?
   A symbol that stands for an entity of the world.

[3] What precisely characterizes cognitivism, according to Varela, is not the commitment to the notion of representation, in a broad and relatively vague sense of the term, but rather the claim that cognition consists in symbol-based operations: "[The] notion of representation is—at least since the demise of behaviorism—relatively uncontroversial. What is controversial is the next step, which is the cognitivist claim that the only way we can account for intelligence and intentionality is to hypothesize that cognition consists of acting on the basis of representations that are physically realized in the form of a symbolic code in the brain or a machine." (Varela, 1991, p. 40).

3) What is The mind?
A system capable of rule-based operations to carry out cognitive processes.

Metaphysical realism is however a philosophical assumption or standpoint, which cannot be proven by cognitivism: after all, by cognitivist lights, we only ever have access to our mental representations of the external world. Furthermore, in this setting it isn't clear how the correspondence between the entity and the representation is grounded. There is no possible way to step outside our mental representations and observe the real world and the way it relates to our representations of it in the same way the implementer of a Turing machine knows what symbol relates to what entity.

As a mathematical model capable – after training – to solve tasks without relying on the assignment of symbols (except for input and output nodes), the ANN is of special interest to cognitivism. Rather than questioning her conceptual framework – given that computation without representation suddenly seems possible –, the convinced cognitivist supposes that ANNs do in fact rely on internal symbols to solve tasks. Henceforth, she obtains a model that does not need symbol assignment as it *learns* them. Following this position, ANNs offer cognitivism's missing element and can be thought of as some sort of elaborate Turing machine that self-adapts to obtain symbol-like representations of some external properties. They are thought to simultaneously learn how to *represent* important properties – *for* themselves (as if they were the implementer of the Turing machine) – and how to use these representations to solve their task. This stance however relies on the supposition that ANNs do indeed learn Turing-like symbols[4]. Furthermore, it is not clear if cognitivism's initial problem is solved as even by relying on the ANN it is not clear how our representations relate to the external world.

In today's cognitive sciences and deep learning literature, this idea of a self-organizing Turing machine is frequently taken up in a new version of cognitivism: neuro-representationalism (NR). In NR, cognitivism's metaphysical realism takes the specific form of a strong scientific realism: the external entities that exist independently of us become the theoretical entities of modern natural sciences. Additionally the cognitivist-interpreted ANN becomes a model of the brain; in other words, the brain is thought to learn Turing-like mental representations of the world. Finally, NR takes cognitivism's representationalism to its extreme by making a claim about our conscious experience: we experience a brain generated model of these entities (Churchland & Sejnowski, 1990; Mrowca, et al., 2018; Sitzmann, et al., 2020); for criticism see (Zahavi, 2018; Hipólito, 2022). Frith puts it in slogan form: "my perception is not of the world, but of my brain's model of the world" (Frith, 2007). According to NR, agents do not directly perceive a photon and its wavelength but only a mental representation of it – in this case, a certain color (Metzinger, 2009). NR too can be systematized around three major questions:

[4] At this point the reader might argue that the fact that ANNs can be implemented on Turing machines does show that they do rely on symbols. The reason this argument fails is that, in such a case, the numbers on the tape of the Turing machine don't explicitly stand for anything, since it was not the implementer who chose what they stand for.

1) What is The World?
   The physical world, composed of the scientific entities discovered by science.
2) What is A Representation?
   A symbol learned by the brain that encodes an external entity of the world.
3) What is The mind?
   Composed of the cognitive processes responsible for mental phenomena – which consist in neural computations on the basis of the representations – and of the content of these phenomena, consciousness – which is given only by the representations.

NR's concept of representation is pervasive in the DL field. In LeCun et al.'s paper (2015), we find a common cognitivist interpretation of ANNs from the first paragraph, describing them as machines that can "be fed with raw data" and "automatically discover the *representations* needed for detection or classification" (LeCun, et al., 2015). Following this view, the first layers of the ANN operate a *feature extraction*, computing relevant features, or representations, while the last layers adequately combine these – carrying out some form of "reasoning". It is also common to consider the layers as extracting features hierarchically, where the first layers compute "low-level" representations – such as edges in an image – and subsequent layers compute "high-level" representations – such as larger motifs in an image (LeCun, et al., 2015). An analogous interpretation can be found in computational neurosciences, when describing, for example, early vision as extracting low-level representations on the basis of which higher-level reasoning is carried out in the brain (Buduma, et al., 2022) Actually, the frequent implicit presence of cognitivism in DL research is revealed in the specific form that it takes, which is NR (Silver, 2015; Buduma, et al., 2022; LeCun, 2022).

A fundamental condition for the acceptance of the construct of NR is the existence of a pre-given world. Reenacting such an agent/world setting corresponds to a large field of DL called deep reinforcement learning (DRL) (Mnih, et al., 2015; Silver, et al., 2016; Eppe, et al., 2022; Wang, et al., 2022). DRL trains an ANN, considered as an agent, to select the right actions based on the observations (or states) of an external environment in order to maximize potential reward (Fig. 1). In this setting the ANN is typically thought to learn an internal *world model* that *represents* the dynamics of the environment and allows it to predict, reason, and plan (Ha & Schmidhuber, 2018;

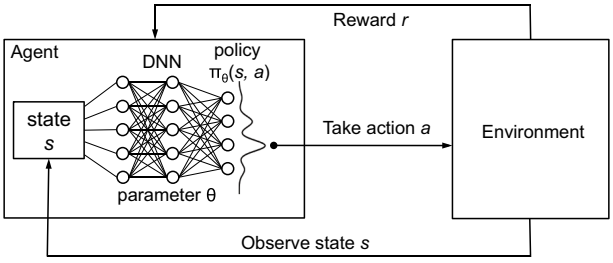

Fig. 1 Model-free deep reinforcement learning. The ANN, or DNN (deep neural network), learns to pick an action based on observations of an external environment, in order to maximize reward. Taken from (Mao, et al., 2016)

Goyal, et al., 2022; Mazzaglia, et al., 2022; Driess et al. 2022). The concept of "world model" is very important among researchers: it is often thought to be the key element to human-level intelligence (Matsuo, et al., 2022). While, in the literature the term "world model" can refer to different things, in the context of DL it means that the ANN does not rely directly on the external environment but operates on the basis of its "world model". Once again, the motivation comes from the cognitivist assumption that an agent does not have direct access to the world and therefore has to "wonder" and "interact" with it through the mediation of internal representations. If this internal character is taken as far as making the world model the realm of perceived things, we get the key ingredient for NR: conscious experience, then, is (limited to) our model of the world. Subsequently, when the ANN learns to extract features from its environment, it is thought to be analogous to the brain learning to "come up" with internal representations of the hidden physical world. By this reasoning, DRL becomes a direct mathematical formalization of NR. In the Machine Learning research community, DRL – a growing field in neuroscience (Botvinick, et al., 2020) – is almost exclusively interpreted in this NR setting – in fact, the brain-in-the-world analogy is typically used to introduce the field of DRL (Fig. 2).

In short, although DL stems from connectionism, a program compelled to reject the challenges that come with assuming (Turing) symbol computation mechanisms on the neurobiological level, the majority of research today carried out in DL still incorporates the cognitivist toolkit, employing concepts (such as world models, representation, etc.) which, then, returns cognition as reducing to NR.

The DL researcher might however object that the concept of representation is of technical use in numerous applications. Indeed, the term representation is often used as a synonym of the term *embedding*, a low-dimensional feature typically obtained by extracting node activations at a particular layer of a trained ANN. When working with images for example, an engineer can extract the pen-ultimate layer's activations of an ANN trained on a first task and use it as a representation, or embedding, on a downstream task such as image classification (Chen, et al., 2020). Because this procedure works, the DL engineer might argue that ANNs do indeed rely on rep-

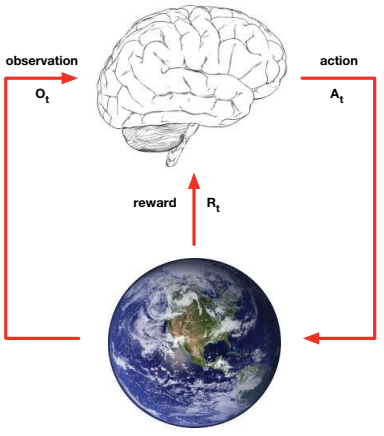

Fig. 2 Figure from deep mind's introduction to reinforcement learning course (Silver, 2015). To introduce deep reinforcement learning, the brain-in-world analogy is used to explain the situation of the ANN in its environment

resentations. This line of thought can be challenged. First of all, these embeddings correspond to one given datum (one precise image for example), not a certain general property of the data. Secondly, they are representations that the engineer chooses for downstream tasks, they are representations for the scientist; they do not stand one-to-one for the datum – in fact, no perfect inverse mapping, that would allow retrieving the datum from an embedding alone, is technically possible[5]. The technical utility of embeddings only shows that patterns of activations at a particular level contain relevant material – that might be, in essence, non-decomposable – to solve a number of tasks. Should this reasoning hold, embeddings – just as the activations of an ANN – are hardly representations in the sense of Turing machine symbols.

In conclusion, little technical or empirical reasons plead in favor of the employment of Turing-like representations *for the ANN*. While in some cases it can be established that some layers are more sensitive to particular patterns (for example edges in an image), the weights and activations of ANNs are notoriously hard to interpret (Zhang, et al., 2021). Even when some interpretation is possible (Olah, 2015; Goh, et al., 2021), the particular patterns could simply represent particular properties of some data *for the engineer/scientist* that interprets them given a certain task (Boge, 2022). In any case, if some form of interpretation is possible decomposition into symbol-based operations seems nowhere near. This is why we employ the concept of ANN "non-decomposability" – as in non-decomposable into symbol-based operations – rather than vaguer "uninterpretability" that is sometimes used in DL literature.

Saying that ANNs are non-decomposable is spelled out precisely by saying that they are not elementwise representational (ER). ER would be obtained if "all model elements (variables, relations, and parameters) represent an element in the phenomenon (components, dependencies, properties)" (Freiesleben, et al., 2022). Such a strong form of representationalism does not seem to hold for ANNs even for quite simple setups (Freiesleben, et al., 2022). Nevertheless, we argue that the ideal of decomposing ANN operations all the way down to elements that represent external properties is still very much present in the field and that it stems from a particular philosophical framework, that is cognitivism. Motivated by the functioning of the TM, it seeks to identify symbol-like representations in ANNs and interpretable rules according to which the representations are manipulated. Possibly because of the technical difficulty (or impossibility) to realize this ideal, some authors argue that ANNs' operations correspond to a sub-symbolic level that realizes symbolic operations on a higher level (Varela, 1991: p 100). It seems unclear however how an ANN can carry out higher-level symbol based operations without relying on any symbol-like operations internally.

In the next sections we will propose a conceptual framework for DL that doesn't rely on this cognitivist ideal of decomposability. We will do so by considering deep learning from phenomenology.

[5] When computing embeddings, the DL engineer wants to reduce the dimensionality of items in a dataset using an ANN. The inversion of this projection is not unique, as infinitely different inputs can generate the same embedding (for example as shown in DeepDream (Mordvintsev, 2015).

3 The Phenomenological Critique of Cognitivism: Implications for Deep Learning

As seen in the previous section, NR takes a robust scientific realist stance, i.e. existence of an external pre-given objective world (that can be described by science and represented by cognitive processes); from then on, the question of consciousness becomes: how does it arise from natural processes? This is notoriously difficult and referred to as the hard problem of consciousness (Chalmers, 1995).

Phenomenology flips the problem around[6]. Instead of positing the existence of an outside world and questioning the emergence of consciousness, phenomenology seeks to describe how the world appears to us in lived phenomena (Merleau-Ponty, 1945). It puts on hold any question regarding the existence of the external world (i.e., whether something is *really* out there or if we are simply hallucinating our reality), its ontology (i.e., the types of things that really exist), and the mind-world relationship (idealism, realism, etc.). Husserl (1931, § 32: 59–60) calls this bracketing of judgment "*epoché*", which aims at neutralizing what he calls the "natural attitude". This attitude is characterized by a common-sense belief in the reality of external, discrete, ordinary objects. Merleau-Ponty (2012: p. 69) calls this natural attitude "objective thought" and interprets it as the shared assumption of idealism and realism, and as the unquestioned metaphysics of modern natural sciences[7]. Once every judgment pertaining to the natural attitude/objective thought has been put into brackets, phenomenology takes as a starting point a passive stance with regards to phenomena. That is, it lets things appear as they appear spontaneously to the mind that is directed towards the world without trying to categorize them. The task of phenomenology is *then* to describe the structures of manifestation, producing an understanding of the mind and its interactions with the world that differs strongly from the views exposed in the previous section. One of the most fundamental features of the mind that phenomenology emphasizes is *intentionality*, that is the fact that mental states are *directed towards* something. For example, an episode of perception is always a perception *of* something. Following Husserl (1900; 1931, § 37), we will call the thing towards which the mind is directed the "intentional object", and the conscious mental state directed towards the intentional object an "intentional act".

Phenomenology and NR diverge significantly in the way they describe cognitive processes. For NR, which operates on a clear separation between subject and external world, a cognitive process can be cast as a sequence of symbol-based operations, from perception (sense data inputs) to a particular action (motor output), or storage of a new useful representation. With its bracketing, phenomenology considers cognitive processes from a different point of view where it makes no sense to distinguish

[6] It is impossible for us to offer an all-encompassing account of the tremendously rich phenomenological tradition or to engage with exegetical issues pertaining to the thought of Husserl and other important thinkers in said tradition. We will limit ourselves to highlighting the points that are essential for the approach we want to sketch in these pages. For detail, see (Zahavi, 2008; Gallagher & Zahavi, 2020).

[7] It is important to note that phenomenology is not per se incompatible with a moderate scientific realism which claims that natural sciences discovers real objective features of the world and produces true statements about it (see, for example, Dreyfus, 1992), although it rejects the idea that the scientific image of the world is a complete and exhaustive descriptive-explanatory framework.

an external entity from our representation of it; there are simply intentional objects that appear to me: consciousness and the world are given in one stroke. Therefore, cognitive processes are not considered as an algorithmic processing of perceptual inputs, but rather as habits that underlie and structure our lived experience, "that simultaneously [delimit] our field of vision and our field of action" (Merleau-Ponty, 1945, p. 153).

Merleau-Ponty (1945, p. 143) offers a useful illustration of the problem posed to those conceiving of perception as indirect: the blind man's cane. He depicts and opposes the position of the intellectualist (which seems identical to NR's position), according to which the blind man infers the shape of external objects in two steps: first, by deducing the cane's position given its pressure on the hand and then, by inferring the shape given this position (Merleau-Ponty, 1945, p. 153). He argues that once that the blindman gets used to the cane, once he has it "in hand", the habit precisely "relieves [him] of this very task" (Merleau-Ponty, 1945, p. 153). The cane becomes "an instrument *with* which he perceives", its tip is "transformed into a sensitive zone", expanding his perceived world (Merleau-Ponty, 1945, p. 154). The acquired cane-sensing skill is simultaneously a perceptual habit and a motor habit (there is no perception without movement of the cane) that structures conscious experience. It grounds an "organic relation between the subject and the world" that does not rely on symbol-like representations (Merleau-Ponty, 1945, p. 154).

Could this different perspective on cognition open up a different way to interpret computational models? Dreyfus, drawing upon Heidegger's phenomenology, considers cognitive processes as acquired habits (and consequently opposes cognitivism's representationalism). He argues that we do not acquire skills by storing representations but by a gradual refinement of our perception that offers new solicitations in given situations in the world; therefore, "the best model of the world is the world itself" (Dreyfus, 2007). He, for example, rejects the existence of an internal map: "what we have learned from our experience of finding our way around in a city is 'sedimented' in how that city *looks* to us" (Dreyfus, 2007, p. 1144). Dreyfus insists that basic cognitive processes (he gives the examples of driving a car or playing chess) do not rely explicitly on symbols but are just the result of a gradual adaptation – they are representation-less (Dreyfus, 2002). Therefore, he sees the advent of ANNs as a strong blow against cognitivism's commitment to representations as they "provide a model of how the past can affect present perception and action without the brain needing to store specific memories at all" (Dreyfus, 2002, p. 374). From the phenomenological standpoint, ANNs seeming non-decomposabilty is particularly interesting because it evokes the opacity of our implicit habits (i.e. that do not seem to rely on representations).

However, the use of the term "brain" in Dreyfus's previous citation marks an important conceptual shift that shouldn't go unnoticed. Important observations that were acquired from the phenomenological source are mapped onto the functioning of the brain – a system understood in terms of and by scientific investigation that belongs to the neurophysiological source. Dreyfus's implicit supposition of an overlap between the phenomenological and the neurophysiological is in tension with phenomenology's initial ambition to put on hold the question of existence of external objects. And this supposition can be taken a step further by trying to reduce the phe-

nomenological to the natural – shifting from cognition explained from first-person perspective to cognition explained from third-person perspective.

This *naturalization* of phenomenology is constitutive of Varela's formulation of enactivism, proposed as an alternative to cognitivism (Varela, et al., 1991). His enactivism retains all major principles uncovered by the phenomenological source and uses them to think the coupling processes of an agent with its environment, considered in a naturalistic setting. It deems that, through sensorimotor activity, organisms become structurally coupled to their environment which allows them to *enact* a world. The "organic relation" between consciousness and world from first-person perspective becomes a "structural coupling" between organism and environment (Varela, et al., 1991, p. 206). When considering the relation between mind and computational models, this enactivism is interested in robots that can navigate autonomously in the external world; Varela et al. (1991) for example turn towards Brooks's finite state machines (Brooks, 1991). This brings it closer to the understanding of cognitive processes as a way to interact with the external world, rather than understanding them as habits that underlie our lived experience. Because enactivism aims to bridge phenomenology and neurophysiology, it operates upon a third-person separation between agent and environment; phenomenology doesn't rely on this dissociation and considers experiences (that are structured by habits) in which the I and the world are intertwined. We therefore insist on the necessity of a clear epistemic separation between the phenomenological source and the neurophysiological source (in the tradition of Husserl's *epoché*) to approach the question of the relationship between cognition and computation. This distinction opens up the possibility of establishing parallels between cognitive processes seen as *habits* and computational models in a way that is more faithful to phenomenology's initial project. Parallels that will be useful to develop a representation-less toolkit for DL.

4 Computational Phenomenology and Deep Learning

In the previous section, we have highlighted the possibility to consider computational models from phenomenology without reducing it to structures uncovered by third-person sciences, such as by neurophysiology[8]. The first such method that distinguished phenomenology, neurophysiology and mathematical/computational modeling is neurophenomenology (Varela, 1996). Neurophenomenology aims to establish "reciprocal constraints" between first-person data from phenomenology and measured data of physiological processes, in order to allow a dialogue, or a "circulation", between the internal (phenomenological) and external (scientific) accounts of a given cognitive process (Varela, 1996, p. 343); a third component is then used to provide a "neutral ground" between "these two kinds of accounts: formal (or computational) models (Lutz & Thompson, 2003). Some approaches specifically focus on math-

[8] Note that mathematical formalizations tend to be seen only (possibly wrongly) as tools to describe objective natural processes. Phenomenology has therefore traditionally been skeptical of any attempts of mathematical formalization. Husserl for example, famously described Galileo as "at once a discovering and a concealing genius" because his mathematization of natural phenomena ends up covering up its origin: lived phenomena (Husserl, 1936, p. 53).

ematically formalizing the structures of lived experience directly, referred to as the "CREA proposal" (Petitot & Smith, 1996; Petitot, 1999). However their obtained mathematical formalizations correspond to a "physical behavior" from which the "qualitative structure of a phenomenon" *emerges* (Petitot & Smith, 1996, p. 241); the computational model describes neurophysiological processes from which phenomenological events emerge (similar to connectionism).

More recently, Yoshimi (2011) proposed to associate phenomenological structures with neuro-computational structures, considering connectionist models. The first approach to propose computational models of the structures of lived experience *without* necessarily assuming that they belong to the neurophysiological is Ramstead et al.'s (2022) computational phenomenology. This active inference-inspired approach seeks to build generative models that explain data from "our first-person phenomenology *itself*" (Ramstead, et al., 2022). In a second step, such models (of perception, or of action for example) can also be used in simulations (Sandved-Smith, et al., 2021).

What all these approaches have not accounted for fully is the "trainability" aspect: computational models can now be *trained*, and therefore generate their own kind of data, in the same way as the neurophysiological and the phenomenological can. In our computational phenomenology account, we propose to use the term *source* to designate a distinct level of inquiry that can provide its own type of data[9]. Following the three fields of knowledge of Lutz and Thompson (2003), we therefore propose to distinguish three different *sources* to explore cognition: *the neurophysiological source*, *the phenomenological source* and *the computational source* – insisting that computational models are not simply tools to formalize data from other fields of knowledge but that they can generate their own.

In fact, neuroscience researchers have long ago recognized the computational as a source of exploration for cognition. Notably, the advent of this new source has strengthened the tendency of neuroscientific research to cast aside the phenomenological source by implicitly assuming a close correspondence between physiological processes and conscious experience. In DL-based computational neuroscience, numerous research completely dismisses the phenomenological source and compares ANN processes to brain processes (comparison which does seem quite natural given DL's initial structural inspiration being the brain); examples include Yamins' paradigmatic research on vision in brains and in ANNs (Yamins & DiCarlo, 2016), other recent investigations (Kumar, et al., 2022; Millet, et al., 2022) and even proposed research programs (Doerig, et al., 2022; Cohen, et al., 2022).

To restore the phenomenological source and give it an epistemic role in DL research, we propose a new formulation of *computational phenomenology* (CP), defining it as an epistemic dialogue between the phenomenological source and computational source (in an analogous way to neurophenomenology). Specifically, we consider circulation of knowledge between phenomenological descriptions of a given cognitive process (and its corresponding habit) and the mechanisms of trained com-

[9] A source is therefore defined epistemologically and not ontologically: it doesn't designate a level of reality but a set of methods from which we can obtain a specific type of data. We owe the term to Bitbol (2006).

putational models on a corresponding task. In line with the phenomenological tradition, we bracket the question of the existence of the external world and provisionally let the two sources free of any constraint or import coming from the neurophysiological source. Therefore CP doesn't help itself to the third-person separation between agent and environment (which neurophenomenology as well as enactivism rely on) and considers cognitive processes, and corresponding habits, in which consciousness and the world are intertwined. *CP's goal is to uncover structural invariants and guiding constraints between phenomenology and computational modeling*.[10] However, it also provides a ground to develop a new framework to think computational models different from the existing ones (cognitivism, connectionism or NR).[11] Figure 3 illustrates the three sources of exploration of cognition and the subsequent inquiries they can generate.

Once the computational is recast as a *source* (because it generates data), it makes sense to select the models that yield the most impressive results, that are best at solving cognitive tasks. Therefore it is natural for CP to turn towards DL. CP is thus not simply the application of a vocabulary derived from phenomenology to DL. It is indeed a dialogue between two sources and phenomenologists could just as well let their practice be informed by the concepts and formal tools developed in deep learning as the reverse. In the next sections, we will engage in such a dialogue between phenomenology and deep learning systematically, covering important dimensions of cognition: perception, action, imagination and language. We will see that ANNs non-decomposaibilty is particularly interesting when considering phenomenological descriptions. However, we will not limit our investigation to highlight a shared opacity in DL and in phenomenology. We will rely on common DL interpretation techniques[12] to highlight similarities between specific cognitive processes, cast as habits, and ANNs trained on an analogous task. This investigation will notably allow us to redefine the term of *representation*. Specifically, we abandon the idea of symbol-like representations underlying our cognition and will characterize the decomposable phenomenal content of our cognitive processes – that means, our conscious representations. In what follows, we offer clear directions within our account of computational phenomenology to its employment within specific DL/cognition areas: perception (4.1), imagination (4.2), and language (4.3).

[10] Furthermore, we must emphasize that the question of the possibility of artificial consciousness or phenomenal experience in ANNs is outside the scope of CP. The fact that there are structural invariants between phenomenology and computational modeling for a given cognitive task does not imply that all systems exhibiting such structures possess a phenomenal experience. One of the goals of CP is to establish which type(s) of mathematical modeling is most adequate to match certain structures of experience, but not to establish a metaphysics of mind that would, for example, define the physical structure that an entity must have in order to be conscious.

[11] Our CP approach can be considered a "cousin" of Ramstead et al.'s active inference formulation (2022). The main difference being that Ramstead' et al.s CP mathematizes the underlying structures of experience directly, whereas our version identifies common mechanisms between AI systems and corresponding first-person experiences of cognitive processes in the perspective to formulate an alternative to cognitivism.

[12] For a review of such interpretation techniques see (Räuker, et al., 2023).

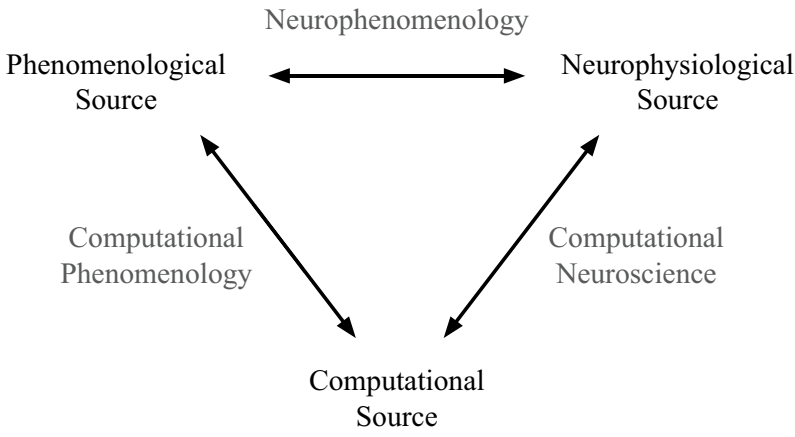

Fig. 3 The three sources of exploration of cognition. Modern computational sciences promote a circulation of knowledge between neuroscience and computational sciences ; in consequence they tend to exclude the phenomenological aspect of cognition. In response, we propose computational phenomenology, an epistemic dialogue between the phenomenological and the computational, in a similar vein than neurophenomenology, a dialogue between the neurophysiological and the phenomenological

4.1 Merleau-pontian Perception: A New Setting in which to Consider Learning

Merleau-Ponty's phenomenology gives a central role to the body (Merleau-Ponty, 1945). In his theoretical framework, the functional unit that allows the imbrication of consciousness and world through action and perception is the *corps propre* (one's own body). The *corps propre* is not the body considered as an object, but the body I live as, and through which I have a world. It is the body I discover by investigating the underlying structures of the first-person perspective; as opposed to the biological body discovered by "objective thought" that is "unaware of the subject of perception" (Merleau-Ponty, 1945, p. 214). It does not rely on the concept of brain, as the brain cannot be identified as the origin of perception from the lived phenomena themselves just as "it would be absurd to say that I see with my eyes or that I hear with my ears" (Merleau-Ponty, 1945, p. 214). When it comes to explaining cognition, the biological body (including the brain) is the functional unit of the neurophysiological source whereas the *corps propre* is the functional unit of the phenomenological source. Considering perception by modeling the former as an ANN tends to fall into NR. An alternative is given by investigating the similarities between the latter and DL.

It turns out the ANN is a good candidate in modeling the way the *corps propre* adjusts its grip on the world. The *corps propre* enables learning from past experiences by *sedimentation* into *habits*. Merleau-Ponty uses this term to stress that experiences aren't stored separately, as symbols in a Turing machine or entries on a hard disk, but rather consolidate into new habits in an analogous way solid material settles (sediments) at the bottom of a liquid. The seemingly uninterpretable way the particles settle resembles the adjustment of weights during the training of an ANN. The sedimented habits form a "contracted knowledge" that isn't "an inert mass at the foundation of our consciousness" but is "taken up" in every "new movement" towards the

world (Merleau-Ponty, 1945, p. 132) in the same way that at inference, all the weights of an ANN are mobilized to propagate the content of each input. In accordance with ANNs that do not store all the examples in the form of representations but rather benefit from them by weight adjustment, the *corps propre*'s intentional acts are not based on a superposition, but a sedimentation of experiences. Intentional acts allowed by the *corps propre* include the *perceptual synthesis*: the detachment of a privileged object from an indifferent background. As such our perceptions are grounded on a perceptual tradition, that contracts the history of previous experiences in a way that remains wholly opaque to the subject of perception – opacity, or non-decomposability, that is also characteristic of ANNs.

That perception is grounded on the sedimentation of past experiences means that what we learn is directly transcribed in the way the world shows up to us. The adjustment of perception corresponds to the search for an *optimal grip* on the lived world that allows confident action. This means that the features of the world that appear to us do not correspond to some encoded properties of external realities. This finding turns some of NR's most elementary assumptions completely on their head: the space that I experience in perception, for example, is now conceived as the result of my grip on the world. The orientation of the whole of our perception – what we consider as being "top", "bottom", "left" and "right" – at any given time, translates a certain equilibrium I reached in my lived world, rather than the encoding of some universally given directions.

But how does this redefinition of perception – as a dynamically adjusted foundation of our experiences rather than a faithful reconstruction of external entities – allow a dialogue with DL models? Let's consider two related cases where perception is perturbed by rotation of our visual field. When I look at an upside-down face for some time, it becomes "monstrous": I see a "pointed and hairless head" with a "blood-red orifice" on its forehead and "two moving eyeballs" where the mouth should be (Merleau-Ponty, 1945, p. 263). My lack of interaction with inverted faces translates into an unstable *grip*. An ANN trained to recognize faces that would be presented to it systematically the "right side up", would fail if they were suddenly turned around. In such cases, the neural network is said to have difficulty generalizing, i.e., adapting to a new type of data. Analogously to the *corps propre*, an ANN is constantly trying to adjust its "grip" on the data, and doesn't rely on a translation into symbols – that, in this case, would have allowed it to instantly revert the inflicted rotation.

The *corps propre* can also readapt in cases of spatial level shifting. In one of Wertheimer's experiments, a subject observing his room for a few minutes through a mirror that tilts it by 45 degrees, suddenly sees his visual spatial level shift when he projects himself into this new setting. The "spectacle" offered by the tilted room is a call to a new "virtual" *corps propre*. That is, the body with "the legs and arms that it would take to walk and act in" the room, to open the cupboard or sit at the table (Merleau-Ponty, 1945, p. 289). At that point perception adjusts in a way to guide actions in this new setting: the spatial level shifts. What exactly triggers this tilting? Before the shift, the orientation of the (intentional) objects in the mirror is not natural and does not allow the usual interaction with them. It is therefore the objects – and particularly faces, according to Merleau-Ponty – that are the sign of a tilted visual

field and serve as anchor points to switch to a new oriented phenomenal space in which they are the "right side up". It turns out an ANN learns a similar "trick" to identify rotated images: *RotNet*, trained to identify the rotation angle (0, 90, 180, or 270 degrees) of rotated images, also relies on the orientation of objects and faces present in the scene (Gidaris, et al., 2018). Indeed, attention maps – which overlay the input image with network activations – reveal a high attention attributed to faces in rotation angle classification (Fig. 4). The orientation of the objects themselves, which allows them to serve as anchor points, is given by a "perceptual itinerary", a learned order in which I visit its important features. *DeepFace* (Taigman, et al., 2014), a neural network that learns to identify a specific person from their face, also relies on certain learned key points as shown by higher activations at pixels corresponding to the eyes and mouth (Fig. 5).[13]

We have first highlighted similarities between the *corps propre*'s grip adjustment and the training of an ANN; we have then shown a case where an ANN and corps propre learn a similar "trick". But what do these correspondences actually show? First, it is interesting to note that simple mathematical functions fitted with a simple learning rule (in this case, gradient descent) end up relying on similar tricks than the *corps propre* to solve analogous tasks. Furthermore, they show how computational learning can be conceptualized in the lived world, in conscious experience, rather than with respect to an inaccessible external world. Indeed, the input to the spatial level adjustment process are not some photons or other scientific entities but our previous lived experience (which calls for an adjustment if it does not allow an optimal grip). Thinking these kinds of processes in terms of ANNs paves the way for a new framework to conceive DL, in which learning happens in the lived world. Furthermore, ANNs' non-decomposability is good news in this setting because it opens up the idea that some processes could, by nature, be non-decomposable into operations on symbols – in the same way adjustment processes are *opaque* to the subject of perception. They allow us to interpret cognition from the phenomenological source without relying on unconscious representations. As for conscious representations (phenomenal contents), we have seen with the case of the orientation of space or of intentional objects, that they are not the encoding of some external entity but the result of an optimal grip on the world. The training of ANNs can therefore, by analogy, be thought of as the gradual constitution of an artificial optimal grip, which means they converge towards an equilibrium given a task and some data, without having to learn specific fine-grained representations of properties[14]. We will now seek to further characterize conscious representations in the two following sections.

[13] However, these key points are not visited in a certain order. The ANNs inference lacks the temporal dimension of our lived perceptual synthesis.

[14] The analogy is limited to this non-representational aspect of optimal grip adjustment, as ANNs typically lack the embodied aspect of the corps propre that places it in a situation of constant self-adjustment. As we only consider particular habits and analogous computational tasks, this shared non-decomposability is sufficient.

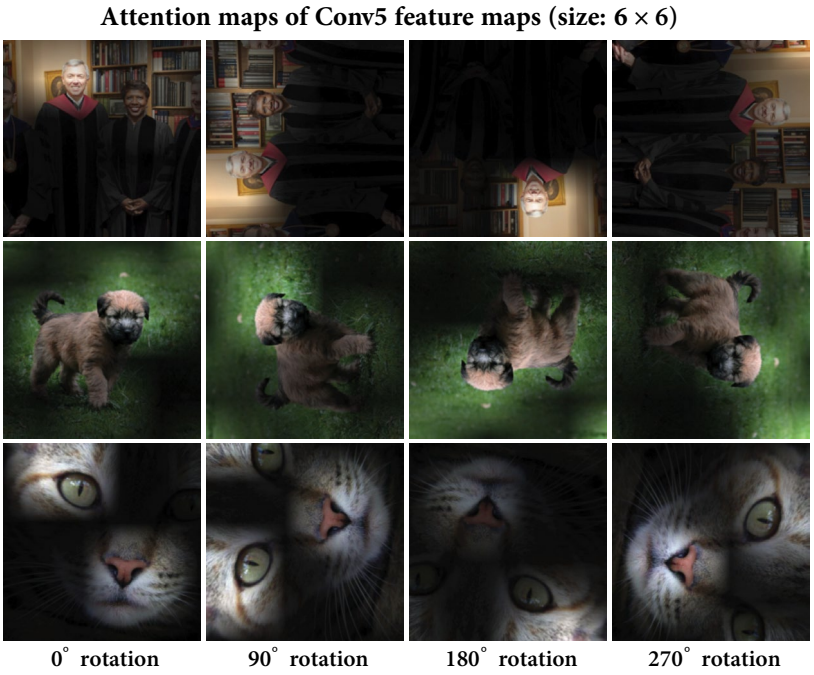

Fig. 4 *Attention maps* of the artificial neural network *Rotnet* that learns to recognize the angle of rotation of an image (Gidaris, 2018)

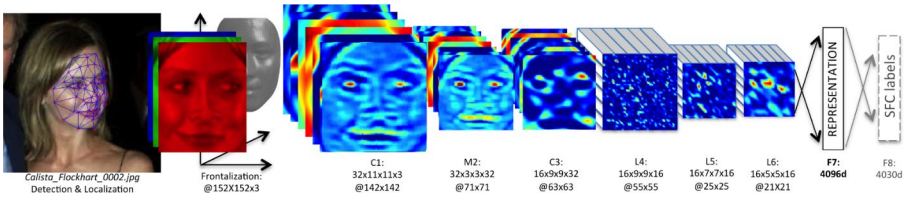

Fig. 5 Architecture of the *DeepFace* model and activations corresponding to a given input. First blue square from the left is the most telling (Taigman, 2014)

4.2 Sartrean Imagination: Conscious Re-presentation of Sedimented Experiences

If ANNs are thought to model processes of the *corps propre*, their training is recast as a *sedimentation* of past experiences in order to obtain an artificial *optimal grip* on the *lived world* by refining the *perceptual synthesis* which determines the way things appear to us. But how can the world be "the best model of itself" when I *imagine* something? Do we not store some representations of things to be able to later form mental images of them? Indeed, cognitivism might turn to this form of explanation relying on the Turing-machine's functioning, positing that we save some representa-

tions to be able to later access them. This explanation corresponds to Hume's understanding of mental images as faint copies of past perceptions. In *The imaginary*, Sartre rejects Hume's copy principle because it falls under what he calls the "illusion of immanence". He holds that mental images are not copies *contained* in consciousness and recasts them as *imaging consciousnesses* (Sartre, 2004). If Hume's copy principle is best explained with the Turing-machine, we will show that Sartre's *imaging consciousness* better resembles the cascade of activations of an ANN.

Sartre rejects the idea of the mental image as a copy; in fact, he considers it as incompatible with the dynamic nature of consciousness: indeed, it would be "impossible to slip these material portraits into a conscious synthetic structure without destroying the structure, cutting the contacts, stopping the current, breaking the continuity." (Sartre, 2004, p. 6). The act of imagining something cannot rely on a symbol, or any object it would be heterogeneous to; rather, the act of imagining and the mental image are one. "The majority of psychologists [mistakenly] think that they find the image in taking a cross-section through the current of consciousness", Sartre says (Sartre, 2004, p. 15); this error is repeated by anyone trying to interpret DL using the cognitivism framework by trying to find representations in the "current" of successful activations of an ANN inference. Therefore, we should rethink the mental image – that is an *imaging consciousness* – as the complete inference of an ANN.

What does it mean for a mental image to be an imaging consciousness? And why cannot it be a representation that would be retrieved by consciousness? Just like the *perceptual synthesis*, the *imaging consciousness* is an intentional act. And in both processes, the object is aimed at as a corporeal object while never *entering* my consciousness. Indeed, when I perceive the chair, it would be absurd to say that the chair *enters* into my consciousness; in the same way an *imagining consciousness* does not rely on a copy of a chair that would be *in* my consciousness. Both intentional acts do not rely on separable symbols. However, even if it relies on a perceptual tradition, the *perceptual synthesis* "encounters" the object it aims at (Sartre, 2004, p. 7). This does not seem to be the case for the *imagining consciousness*. On what knowledge can it rely? How can it aim at the sensitive elements of objects it first encountered in perception?

To picture something as possessing certain sensible qualities, to aim at my friend as "blond, tall, with a snub or aquiline nose, etc.", my *imaging consciousness* "must aim through a certain layer of consciousness that we can call the layer of knowledge" (Sartre, 2004, p. 57). From Sartre's descriptions, we can formulate a hypothesis according to which the *imaging consciousness* re-employs certain tracks of the *perceptual synthesis*. In DL mechanisms, this could mean that the imaging consciousness uses the sedimented weights of processes belonging to the *perceptual synthesis* (such as the process that allows the orientation of an object). In fact, there exists a technique in DL – called *DeepDream* – that allows one to generate images from a trained image classification network: for example, by tweaking random noise in order to maximize the activation of a final output node corresponding to one particular label (Mordvintsev, et al., 2015) (Fig. 5). Recently, (Fei, et al., 2022) have employed a similar technique to visualize what they call the "imagination" of a model trained

to bring closer an image and a matching textual description in a higher-level feature space. In a similar way, my *imaging consciousness* could try to strongly re-activate the processes that allow me to recognize (in perception) my friend as being "blond, tall, with a snub or aquiline nose, etc.". This, of course, would be only one possible way to do it; the general idea we want to convey is that past perceptual experiences might *sediment* into perceptual *habits*, and that in a second phase some structures, or some "regions", of these habits might be used, or *re-activated*, by *imaging consciousnesses* (which are themselves are particular type of habits). As such, an imagined object isn't really red because we stored its color in symbol form, it is red because we employ a certain red-making process (that relies on the process used to recognize objects as red in perception): we imagine *redly*.

Once the mental image recast as an *imaging consciousness*, Sartre establishes two characteristics of the aimed object in the way it appears to us. These also apply, analogously, to images generated by DL. Mental images are given as an "undifferentiated whole", meaning that they can possess qualitative properties but be quantitatively undefined: it is possible for example to have a mental image of the Pantheon but to be unable to count the columns – it can have an undetermined number of them (Sartre, 2004, p. 87). Furthermore, they do not obey the "principle of individuation": when I imagine my friend for example, she can appear simultaneously from the front and the profile. Both the qualitative undefinition and the superposition of various viewpoints can be observed in images generated by *DeepDream* (Fig. 6) or other DL image generations techniques (Fei, et al., 2022; Ramesh, et al., 2021). These properties, or failure cases, of imagination from the phenomenological source can be observed analogously in DL based image generation. The TM, on the other hand, cannot explain an "undifferentiated" Pantheon; it is more in line with Hume's copy principle, where one either has a stored copy of the Pantheon and can retrieve or one does not and cannot output anything. Dedicated DL models better explain borderline cases of mental images considered from the phenomenological source than the TM, cognitivism's go-to model, does.

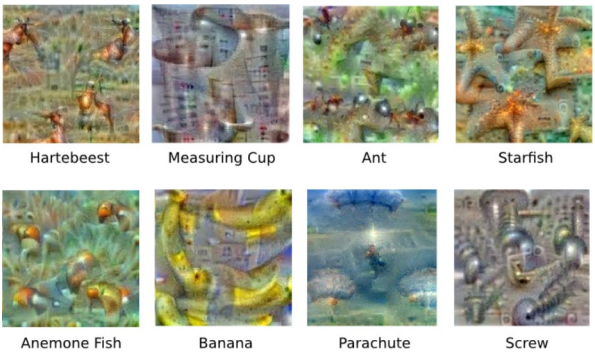

Fig. 6 *DeepDream* images that maximize activations corresponding to particular image labels. (Mordvintsev, 2015)

In conclusion, Sartre's concept of *imaging consciousness* can be modeled by the inference of an ANN taken as a whole process, wherein one does not look for any symbol-like representations. It relies on previously *sedimented* perceptions allowing it to *re-present* (show again) properties aimed at by the *perceptual synthesis* by *re-activating* them. Examples of such properties include the color and even the orientation of an object. We use the term *re-presentation* to stress that previous experiential content is *re-activated*, rather than symbolically represented: we imagine *redly* instead of accessing the symbol red in our internal database[15]. Analogously, the inference of an ANN re-activates particular regions given its weights. If a particular region can be identified to play a specific role in the transformation carried out by the ANN its activation patterns can be seen as a re-presentation. With this understanding of re-presentations, we can drop the idea of symbol-like representations that serve as a mediation between the data and the ANNs operations.

4.3 Language: Concepts as Conscious Re-presentations

Computational phenomenology also has to address a major aspect of cognition that is also an important research domain in deep learning: language. Language and concept-based reasoning heavily rely on individuated units or conscious symbols. These units provide the ability to group and recognize things by subsuming them under the same concept.[16] These concepts can be employed even in the absence of any instance of it, for example in a propositional act. In such a case a concept is linked with a series of sounds, to a word. These "meaningful cores" can also be explored from the phenomenological source, although we must immediately indicate certain limits of the correlation of this source with the computational source. Indeed, the lived world, according to Husserl's conception as well as that of Merleau-Ponty, is an intrinsically intersubjective world (Husserl, 1936; Merleau-Ponty, 1945). A phenomenological approach to language cannot do without this intersubjectivity and the communal character of the acquisition, use, and transmission of language. In this sense, having a language and being with others are inseparable. ANNs do not have features equivalent to the intersubjectivity of the lived world, which imposes limits on the scope of

[15] Our conception of re-presentations shares numerous aspects with Barsalou's perceptual symbols (1999). However, our re-presentations are considered from the first-person perspective (and not from the neurophysiological source). Furthermore, they do not *stand for* any external entity as TM symbol would – accordingly, if one were to follow our terminology, the term of perceptual *symbol* would be misleading.

[16] A question that could be addressed to us would be how the use of symbols is possible for a cognition interpreted from a non-representationalist perspective. The most important point, in our opinion, which allows us to dissolve this doubt, is the following: it is necessary to distinguish the idea that cognition consists of operations on internal representations (representationalism) from the idea that cognition can produce and use symbols and representations. Our position amounts to rejecting the first idea while retaining the second.

computational phenomenology for the study of language. However, some cognitive processes take place at the individual level in the use of language, and it is at this level that the correlation of the two sources must be placed in this context. In the rest of this section, we outline the framework within which the dialogue between the two sources can take place on the question of language.

To begin with Merleau-Ponty, a word is a "phonetic gesture", generated through my body and its vocal cords, that "produces a certain structuring of experience" in the same way than my perceptual habits do (Merleau-Ponty, 1945, p. 199). As they crystallize into words, concepts become communicable: in the same movement "the body opens itself to a new behavior and renders that behavior intelligible to external observers" (Merleau-Ponty, 1945, p. 199). We can consider concepts as *conscious re-presentations* by extending the sartrean conception of imagination.

As we did for perception and imagination, we could consider concepts as being cognitive processes that rely on conceptual *habits*. This thesis can be mapped to ANN mechanisms : the inference of the ANN can stand for the concept as a cognitive process (or *conscious re-presentation*), the training – or sedimentation of past experiences – can stand for the formation of the *conceptual habit*. Similarly to mental images concepts might also reemploy some sedimented tracks of other habits (perceptual or imaginative[17]). Here particular ANNs can be used to reason about different functions of concepts understood as cognitive processes. Panaccio proposes two fundamental roles of the concept: its *representational role* and its *inferential role* (Panaccio, 2011).

The *representational role* refers to a concept's relationship to perception allowing recognition and re-presentation. The concept *red* allows me to perceive objects as red in my lived world and also to re-present *redly* in imagination. As such concepts, and associated words, are structuring lived experience; they are not name tags of external properties. Merleau-Ponty stresses this point by analyzing a study of Gelb and Goldstein where patients with color name amnesia were incapable of grouping objects of the same color (Merleau-Ponty, 1945, p. 197). From DL, models of image classification – where a concept (a word) is associated with an image – would then be of interest to investigate the *representational role* of concepts. In such cases, the task of image recognition sediments into weights which determine the pattern of successive activations at inference (perception) and allow re-activation (imagination). The aspect of experience-structuring can also be observed analogously: a recent DL study, for example, showed that semantic segmentation emerges when training a network to caption images (Xu, et al., 2022).

The *inferential role* refers to a concept's relationship to language. Inherent to my concept of a *cat* is the ability it gives me to draw inferences from it: "a cat is a mammal", "a cat has four legs", etc. Interestingly, I can possess only the inferential dimension of a concept: even if I never saw a platypus, and even if I have no idea what it looks like, I might still know that it is a mammal. In such cases, we first

[17] Sartre for example speaks of "illustrations" and "symbolic schemas", that are two different kinds of mental images that can accompany our recollection of a particular concept (Sartre, 2004, p. 87).

encounter the word and then form the corresponding concept based on the things we learn it implies (Panaccio, 2011, p. 61). To model the *inferential role* of concepts drawing from deep learning, we can turn to DL language modeling (Brown, et al., 2020; Devlin, et al., 2019). Under masked language modeling, sentences are fed word by word to an ANN; however, certain words are randomly masked, and the task of the model is to predict them. Once trained, the ANN allows the generation of highly coherent text. Recent DL research has shown promising results for approaches relying on both of these concept roles (Fei, et al., 2022; Li, et al., 2020; Radford, et al., 2021) – two roles that most of the concepts of everyday use play simultaneously.

This last section on mental language and language allows us to generalize our proposal. In the lived world we can isolate some cognitive processes that structure our experience, for example that give the objects their orientation or allow me to recognize them as falling under the concept of *red*. These processes continuously rely on a *sedimentation* of past experiences, which allows an *optimal action-orienting grip* on the world. *Habits*, resulting from the sedimentation, also allow re-presentation, in the form of imagination or recollection: I can form mental images and I can draw inferences from my concepts (even those I only "encountered" through language and not through direct perception). These re-presentations can be thought of as re-activations of sedimented weights of an ANN. Finally, in all these processes, conscious representation (or concept) and cognitive process are one: cognitive processes are not to be decomposed into symbols and their content is to be thought of as a process and not as an object. Both the mechanism of an ANN (processual) and its seeming non-decomposability support this thesis.

5 A New Toolkit for Deep Learning and A Novel Mathematization of Cognition

Computational phenomenology is a novel approach to work towards a computational interpretation of cognitive processes as described from first-person perspective. Rather than mathematizing cognition considered as emerging from the external world as described by science, it takes a new background of exploration: conscious experience. The CP recipe is the following: (1) place oneself in the lived world, (2) isolate a particular cognitive process and (3) compare it to the mechanism that a computational model uses to solve an analogous task. We believe CP has potential for practical applications in both phenomenology and deep learning. By applying the CP recipe, the phenomenologist could find some clues as to what aspects of a particular habit to investigate; he could for example consider the mechanism of RotNet and test if its tricks are also analogously realized by the corps propre (relying on faces for example). Inversely, the CP recipe could provide the deep learning engineer some new ideas to design DL setups : for example by trying to generate images using the weights of trained recognition networks in an analogous way to the mechanism of sartrean imagination. From phenomenology to deep learning,

one could even hope for some insights to tackle some more fundamental aspects of modern DL – could be investigated: a first-person-perspective credible learning rule (an alternative to *backpropagation*) or a way to make the inference of an ANN more temporal (like the habits of the *corps propre* are)[18]. We hope our theoretical groundwork will enable such transdisciplinary approaches, like there have been many for neurophenomenology.

CP is skeptical about the feasibility of a decomposition of ANNs into operations on symbol-like representations that would encode one, and only one, property of the input domain. However, this does not mean that CP rejects all forms of DL interpretation. For instance, we have seen that attention maps used in DL are analogous to the careful attention of the phenomenologist to her experience, and that techniques such as DeepDream, which aim to maximally activate particular nodes, are akin to phenomenological investigations of imagination to understand perception. Furthermore, even when specific units in ANNs have been found to strongly relate to a particular concept, such as in the work of Bau et al. (2019) and Goh et al. (2021), this need not invalidate CP altogether. First, these findings only demonstrate the existence of some concept-specific neurons and only in a non-exclusive manner (meaning that the identified neurons might also play a role in other aspects of the transformation carried out by the ANN). Which makes them compatible with our position according to which some regions of an ANN can play a particular role in the transformation of inputs into outputs. Second, the ablation techniques that are typically used to reveal these particular neurons also have a counterpart in phenomenology. Indeed, Merleau-Ponty famously considered cases of neuro-psychopathologies where specific brain portions were missing and studied their effects on the patients' lived experience (1945). In general interpretation techniques of DL are limited to probing the ANN rather than decomposing it systematically into rule-based operations; this arguably mirrors the position of the phenomenologists that has to find tricks to investigate habit sedimentation while dealing with the opacity of the corps propre's cognitive processes. In the future, developing CP-inspired interpretation techniques for DL could be another promising avenue for investigation.

Applying CP to DL we have arrived at a conceptual framework that lies in stark opposition to cognitivism and neuro-representationalism. We take as a starting point the lived world rather than the world interpreted in physicalist terms. From this perspective, we reject the existence of unconscious symbols underlying our cognitive processes and propose a theory of conscious re-presentations as re-activations of sedimented habits that are fused to their associated processes. Finally, this perspective also does not allow us to draw a boundary between mind and world, like a naturalist can separate an organism from its environment. Consciousness remains the background of our investigation, and even in the investigated cognitive processes a "subject of the cognition" is not easily found – Merleau-Ponty for example, says that perception is better described by "*one* perceives in me" than "I perceive" (Merleau-

[18] This aspect was raised in footnote 13: the perceptual synthesis (one of the habits of the *corps propre*) is temporal – we constantly adjust what we see in time – whereas the inference of an ANN isn't (the sequentiality of its operations can't be equated to lived temporality).

Ponty, 1945, p. 223). Still a functional unit, the *corps propre*, can be isolated to better explain the way I adjust my grip on the world. This has an interesting consequence: from CP the quest for an autonomous agent, or even an AGI, is not necessarily the ultimate goal, as we rather isolate and study particular cognitive processes (of the *corps propre*). Finally, to summarize:

1) What is The World?
   The lived world, that can be investigated from first-person perspective.
2) What is A Representation?
   A conscious representation is an act that is inextricably linked to its corresponding process. This act is realized by *re-activation* of *habits* obtained by *sedimentation* of past experiences.
3) What is The mind?
   The mind is not delimitable. However, cognitive processes can be isolated and explored through consciousness. All the cognitive processes belong to the *corps propre* and have a common objective: obtaining an *optimal grip* on the world.

Unsurprisingly, this framework that we have obtained by applying CP to DL, is better modeled using the mechanisms of ANNs than those of a Turing Machine. As we mentioned previously (cf. Section 4 above), CP does not involve a simple transfer of vocabulary from phenomenology to deep learning, but rather seeks to establish a new methodological approach aimed at integrating the data provided by the phenomenological and the computational sources. In the context of this integration, we propose to consider the training of ANNs as a way to constitute an artificial *optimal grip* on some data, analogously to the way the *corps propre* adjusts its grip on the *lived world*. Instead of learning to faithfully represent existing properties with the use of symbols, ANNs learn to re-employ particular tracks that are the result of their past experiences. In fact, past experiences are sedimented into *habits*. These *habits* can further be reemployed in other tasks by *re-activation*, allowing *re-presentations*. When thought of as modeling processes of the *corps propre*, the ANN's inference can be seen as an *intentional act* – such as the *perceptual synthesis*, the *imaging consciousness*, or even the use of a *lexical concept*. The fact that DL is non-decomposable is not seen as a weakness in this framework, as conscious representations, or concepts, are inextricably linked to their corresponding cognitive processes. We nevertheless agree that some regions might be responsible for some particular aspects of the transformation carried out by the ANN; while rejecting the idea that it relies on operations on internal representations (for itself). In the end, the DL engineer, or computational neuroscientist, might rely on the CP framework or the NR framework depending on the cognitive process he is considering. If she considers learning as happening in lived experience – as when someone is learning to drive a car or play chess for example – she might rely on the former; if she considers it as happening in the physical world – as with the perception of a certain wavelength for example – she could rely on the latter.

We have proposed a new methodology to move the field of DL and phenomenology forward as well as an alternative conceptual toolkit to consider DL. It considers the computational from lived experience and therefore puts on hold the question of the material basis of consciousness. Therefore, it distinguishes itself from cognitivism and neuro-representationalism; but also from connectionism and some forms of enactivism (like Varela's) that think of consciousness in terms of emergence. Notably, it has led us to a novel type mathematization that cannot be conceived from the scientific-realism-grounded NR position. Indeed, by considering ANNs to model the processes structuring our experience, we oppose Galilean mathematization, criticized by Husserl, that necessarily implies a commitment towards the existence of external objects: for example, of physical bodies to which the law of gravitation can apply. As such, CP opens up a novel form of computational modeling – that "precedes" the laws that govern objects – of the processes by which I learn to perceive, imagine, or think things. Become mathematizable not only the physical laws applying to objects but also the mechanisms that allow me to perceive an (intentional) object as a delimited object distinguished from its background – perception of objects from which I can start to formulate the hypothesis of the existence of an external physical world.

In such cases, and more generally for all processes considered by computational phenomenology, what do the mathematical models tell us? What do they correspond to? Two interpretations are possible. Either we remain faithful to classical phenomenology by keeping the neurophysiological source bracketed. In that case, we consider that we are modeling mechanisms of the spontaneous mind that are neither identical nor reducible to the physical states with which they are correlated, but rather allow the creation of a world around us. Furthermore, qualia are part of the starting point of our investigation and their emergence does not need to be explained: the hard problem of consciousness is thus avoided. Otherwise, we return to naturalism and consider that the cognitive processes involved in the unveiling of the physical world are themselves reducible to physical events[19]. The invariant mechanisms we have isolated by applying computational phenomenology – that rely on reactivating sedimented lived experiences – could be confronted with the neural operations of the brain. Such an informed return to the neurophysiological source could provide an opportunity to surpass cognitivism and neuro-representationalism in modern neuroscience[20].

[19] After all, the order in which we learn to conceive things does not necessarily entail an order of existence. As Sellars puts it: "we must distinguish between primacy in the order of being and primacy in the order of conceiving" (Sellars, 1971, p. 408).

[20] Following this second interpretation, a CP-informed neuroscience would reject the notion of neural representations, following neuroscientists that find it misleading (Freeman & Skarda, 1990; Brette, 2019), having identified it as the well-known map-territory fallacy (confusing models of reality with reality itself) (Korzybski, 1933). It might find common grounds with anti-representationalist theories such as the dynamical systems approach (Van Gelder, 1995; Freeman, 2000; Favela, 2021; Hipólito, 2022) or instrumentalist accounts of the Free Energy Principle (Friston, 2013; Andrews, 2021; van Es, 2021; Parr, et al., 2022).

Funding Open access funding provided by University of Lausanne

References

Andrews, M. (2021). The math is not the territory: Navigating the free energy principle. *Biology & Philosophy*, *36*(3), 1–19.

Ashby, F. G. (2014). *Multidimensional models of perception and cognition*. Psychology Press.

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, *33*, 12449–12460.

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

Bau, D., Zhu, J. Y., Strobelt, H., Zhou, B., Tenenbaum, J. B., Freeman, W. T., & Torralba, A. (2019). GaN dissection: Visualizing and understanding generative adversarial networks. In International Conference on Learning Representations. International Conference on Learning Representations, ICLR.

Bechtel, W., & Abrahamsen, A. (1991). *Connectionism and the mind: An introduction to parallel processing in networks*. Basil Blackwell.

Bitbol, M. (2006). Une science de la conscience équitable. L'actualité de la phénoménologie de Francisco Varela. *Intellectica*, *43*(1), 135–157.

Boge, F. J. (2022). Two dimensions of opacity and the deep learning predicament. *Minds and Machines*, *32*(1), https://doi.org/10.1007/s11023-021-09569-4.

Botvinick, M., Wang, J. X., Dabney, W., Miller, K. J., & Kurth-Nelson, Z. (2020). Deep Reinforcement Learning and Its Neuroscientific Implications. In *Neuron* (Vol. 107, Issue 4). https://doi.org/10.1016/j.neuron.2020.06.014

Brette, R. (2019). Is coding a relevant metaphor for the brain?. *Behavioral and Brain Sciences*, 42.

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

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *2020-December*.

Buckner, C. (2019). Deep learning: A philosophical introduction. *Philosophy Compass*, *14*(10), https://doi.org/10.1111/phc3.12625.

Buduma, N., Buduma, N., & Papa, J. (2022). *Fundamentals of deep learning*. O'Reilly Media, Inc.

Castelvecchi, D. (2016). Can we open the black box of AI? *Nature News*, *538*(7623), 20.

Chalmers, D. J. (1995). Facing up to the hard problem of consciousness. *Journal of Consciousness Studies*, *2*(3).

Chemero, A. (2011). *Radical embodied cognitive science*. MIT press.

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. *37th International Conference on Machine Learning, ICML 2020*, *PartF168147-3*.

Churchland, P. S., & Sejnowski, T. J. (1990). Neural representation and neural computation. *Philosophical Perspectives*, *4*, 343–382.

Cohen, Y., Engel, T. A., Langdon, C., Lindsay, G. W., Ott, T., Peters, M. A., & Ramaswamy, S. (2022). Recent advances at the interface of Neuroscience and Artificial neural networks. *Journal of Neuroscience*, *42*(45), 8514–8523.

Davies, A., Veličković, P., Buesing, L., Blackwell, S., Zheng, D., Tomašev, N., & Kohli, P. (2021). Advancing mathematics by guiding human intuition with AI. *Nature*, *600*(7887), 70–74.
Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019–2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, *1*.
DeVries, P. M., Viégas, F., Wattenberg, M., & Meade, B. J. (2018). Deep learning of aftershock patterns following large earthquakes. *Nature*, *560*(7720), 632–634.
Di Paolo, E., Buhrmann, T., & Barandiaran, X. (2017). *Sensorimotor life: An enactive proposal*. Oxford University Press.
Doerig, A., Sommers, R., Seeliger, K., Richards, B., Ismael, J., Lindsay, G., & Kietzmann, T. C. (2022). The neuroconnectionist research programme. *arXiv preprint arXiv:2209.03718*.
Dreyfus, H. L. (1992). 2. Heidegger's Hermeneutic Realism. *The interpretive turn: Philosophy, Science, Culture* (pp. 25–41). Ithaca, NY: Cornell University Press.
Dreyfus, H. L. (2002). Intelligence without representation - Merleau-Ponty's critique of mental representation. *Phenomenology and the Cognitive Sciences*, *1*(4).
Dreyfus, H. L. (2007). Why heideggerian AI failed and how fixing it would require making it more heideggerian. *Artificial Intelligence*, *171*(18), https://doi.org/10.1016/j.artint.2007.10.012.
Driess, D., Ha, J. S., Toussaint, M., & Tedrake, R. (2022, January). Learning models as functionals of signed-distance fields for manipulation planning. In Conference on Robot Learning (pp. 245–255). PMLR.
Eppe, M., Gumbsch, C., Kerzel, M., Nguyen, P. D., Butz, M. V., & Wermter, S. (2022). Intelligent problem-solving as integrated hierarchical reinforcement learning. *Nature Machine Intelligence*, *4*(1), 11–20.
Favela, L. H. (2021). The dynamical renaissance in neuroscience. *Synthese*, *199*(1), 2103–2127.
Fazi, M. B. (2021). Beyond human: Deep learning, explainability and representation. *Theory Culture & Society*, *38*(7–8), 55–77.
Fei, N., Lu, Z., Gao, Y., Yang, G., Huo, Y., Wen, J., Lu, H., Song, R., Gao, X., Xiang, T., Sun, H., & Wen, J. R. (2022). Towards artificial general intelligence via a multimodal foundation model. *Nature Communications*, *13*(1), 3094. https://doi.org/10.1038/s41467-022-30761-2.
Fodor, J. A. (1983). *The modularity of mind*. MIT press.
Freeman, W. J. (2000). *How brains make up their minds*. Columbia University Press.
Freeman, W. J., & Skarda, C. A. (1990). Representations: Who needs them?.
Freiesleben, T., König, G., Molnar, C., & Tejero-Cantero, A. (2022). Scientific inference with interpretable machine learning: Analyzing models to learn about real-world phenomena. *arXiv preprint arXiv*:2206.05487.
Friston, K. (2013). Life as we know it. *Journal of the Royal Society Interface*, *10*(86), 20130475.
Frith, C. (2007). *Making up the mind: How the brain creates our mental worlds*. Oxford: Blackwell.
Gallagher, S. (2017). *Enactivist interventions: Rethinking the mind*. Oxford University Press.
Gallagher, S., & Zahavi, D. (2020). *The phenomenological mind*. Routledge.
Gidaris, S., Singh, P., & Komodakis, N. (2018). Unsupervised representation learning by predicting image rotations. *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*.
Goh, G., Cammarata, N., Voss, C., Carter, S., Petrov, M., Schubert, L., Radford, A., & Olah, C. (2021). Multimodal neurons in Artificial neural networks. *Distill*, *6*(3), https://doi.org/10.23915/distill.00030.
Goyal, A., & Bengio, Y. (2022). Inductive biases for deep learning of higher-level cognition. Proceedings of the Royal Society A, 478(2266), 20210068.
Ha, D., & Schmidhuber, J. (2018). *World Models*. https://doi.org/10.5281/zenodo.1207631
Hipólito, I. (2022). Cognition without neural representation: Dynamics of a Complex System. *Frontiers in Psychology*, 5472.
Hsu, W. N., Bolte, B., Tsai, Y. H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio Speech and Language Processing*, *29*, 3451–3460.
Husserl, E. ([1900] 2001). *Logical investigations volume 1*. Routledge.
Husserl, E. ([1936] 1970). *The Crisis of European Sciences and Transcendental Phenomenology an introduction to Phenomenological Philosophy*. Northwestern University Press.
Husserl, E. (Ed.). ([1931] 2012). *Ideas: General introduction to pure phenomenology*. Routledge.
Hutto, D. D., & Myin, E. (2012). *Radicalizing enactivism: Basic minds without content*. MIT press.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589.

Korzybski, A. (1933). *Science and sanity: An introduction to non-aristotelian systems and general semantics Lakeville*. Conn.: International Non-aristotelian Library Publishing Co.

Kumar, S., Sumers, T. R., Yamakoshi, T., Goldstein, A., Hasson, U., Norman, K. A., Griffiths, T. L., Hawkins, R. D., & Nastase, S. A. (2022). Reconstructing the cascade of language processing in the brain using the internal computations of a transformer-based language model. *BioRxiv*.

LeCun, Y. (2022). A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Lees, R. B., & Chomsky, N. (1957). *Syntactic Structures Language*, 33(3). https://doi.org/10.2307/411160

Li, G., Duan, N., Fang, Y., Gong, M., & Jiang, D. (2020). Unicoder-VL: A universal encoder for vision and language by cross-modal pre-training. *AAAI 2020–34th AAAI Conference on Artificial Intelligence*. https://doi.org/10.1609/aaai.v34i07.6795

Lutz, A., & Thompson, E. (2003). Neurophenomenology: Integrating subjective experience and Brain Dynamics in the neuroscience of consciousness. *Journal of Consciousness Studies*, *10*, 9–10.

MacKay, D., Shannon, C., & McCarthy, J. (1956). *Automata studies*.

Mao, H., Alizadeh, M., Menache, I., & Kandula, S. (2016). Resource management with deep reinforcement learning. In *Proceedings of the 15th ACM Workshop on Hot Topics in Networks* (pp. 50–56).

Matsuo, Y., LeCun, Y., Sahani, M., Precup, D., Silver, D., Sugiyama, M., Uchibe, E., & Morimoto, J. (2022). Deep learning, reinforcement learning, and world models. *Neural Networks*, *152*, 267–275. https://doi.org/10.1016/j.neunet.2022.03.037.

Mazzaglia, P., Verbelen, T., Çatal, O., & Dhoedt, B. (2022). The Free Energy Principle for Perception and Action: A deep learning perspective. *Entropy*, *24*(2), 301.

McClelland, J. L. (2022). *Capturing advanced human cognitive abilities with deep neural networks*. Trends in Cognitive Sciences.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, *5*(4), https://doi.org/10.1007/BF02478259.

Merleau-Ponty, M., & Landes, D. A. (2012). *Phenomenology of perception*. Routledge. ([1945].

Metzinger, T. (2009). *The ego tunnel*. New York: Basic Books.

Milkowski, M. (2013). *Explaining the computational mind*. Mit Press.

Millet, J., Caucheteux, C., Orhan, P., Boubenec, Y., Gramfort, A., Dunbar, E., Pallier, C., & King, J. R. (2022). Toward a realistic model of speech processing in the brain with self-supervised learning. In *arxiv.org*. https://arxiv.org/abs/2206.01685

Minsky, M. (1961). Steps Toward Artificial Intelligence. In *Proceedings of the IRE* (Vol. 49, Issue 1). https://doi.org/10.1109/JRPROC.1961.287775

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., & Hassabis, D. (2015). Human-level control through deep reinforcement learning. *nature*, *518*(7540), 529–533.

Mordvintsev, A., Olah, C., & Tyka, M. (2015). Inceptionism: Going Deeper into Neural Networks. In *Research Blog*.

Mrowca, D., Zhuang, C., Wang, E., Haber, N., Fei-Fei, L. F., Tenenbaum, J., & Yamins, D. L. (2018). *Flexible neural representation for physics prediction* (p. 31). Advances in neural information processing systems.

Olah, C. (2015). Understanding LSTM Networks. *GITHUB Colah Blog*. http://colah.github.io/posts/2015-08-Understanding-LSTMs/

Panaccio, C. (2011). *Qu'Est-Ce Qu'Un Concept?*https://www.cambridge.org/core/services/aop-cambridge-core/content/view/S0012217312000297

Parr, T., Pezzulo, G., & Friston, K. J. (2022). *Active inference: The free energy principle in mind, brain, and behavior*. MIT Press.

Perconti, P., & Plebe, A. (2020). Deep learning and cognitive science. *Cognition*, *203*, 104365.

Petitot, J. (1999). *Naturalizing phenomenology: Issues in contemporary phenomenology and cognitive science*. Stanford University Press.

Petitot, J., & Smith, B. (1996). Physics and the phenomenal world. *Formal ontology* (pp. 233–253). Dordrecht: Springer.

Piantadosi, S. T. (2021). The computational origin of representation. *Minds and machines*, *31*(1), 1–58.

Poldrack, R. A. (2021). The physics of representation. *Synthese*, *199*(1), 1307–1325.
Putnam, H. (1967). The nature of mental states. *Art mind and religion*, 37–48.
Radford, A., Wook, J., Chris, K., Aditya, H., Gabriel, R., Sandhini, G., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from Natural Language Supervision. *OpenAI*, 47. https://github.com/openai/CLIP
Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., & Sutskever, I. (2021). Zero-Shot Text-to-Image Generation. In *proceedings.mlr.press*. https://github.com/openai/DALL-E
Ramstead, M. J. D., Seth, A. K., Hesp, C., Sandved-Smith, L., Mago, J., Lifshitz, M., Pagnoni, G., Smith, R., Dumas, G., Lutz, A., Friston, K., & Constant, A. (2022). From Generative Models to Generative Passages: A Computational Approach to (Neuro) Phenomenology. *Review of Philosophy and Psychology*. https://doi.org/10.1007/s13164-021-00604-y
Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), https://doi.org/10.1037/h0042519.
Räuker, T., Ho, A., Casper, S., & Hadfield-Menell, D. (2023). Toward transparent AI: A survey on interpreting the inner structures of deep neural networks. *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*. IEEE.
Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning Internal Representations By Error Propagation. In *Cognitive Science* (Vol. 1, Issue V).
Saddler, M. R., Gonzalez, R., & McDermott, J. H. (2021). Deep neural network models reveal interplay of peripheral coding and stimulus statistics in pitch perception. *Nature Communications*, *12*(1), 7278.
Sandved-Smith, L., Hesp, C., Mattout, J., Friston, K., Lutz, A., & Ramstead, M. J. D. (2021). Towards a computational phenomenology of mental action: Modelling meta-awareness and attentional control with deep parametric active inference. *Neuroscience of Consciousness*, *2021*(2), https://doi.org/10.1093/nc/niab018.
Sartre, J. P., Elkaïm-Sartre, A., Webber, J., & Jonathan, M. (2004). *The imaginary: A phenomenological psychology of the imagination*. Routledge.
Schulman, J., Zoph, B., Kim, C., Hilton, J., Menick, J., Weng, J., & Ryder, N. (2022). ChatGPT: Optimizing language models for dialogue.
Sellars, W. (1971). Science, sense impressions, and Sensa: A reply to Cornman. *The Review of Metaphysics*, *24*(3), 391–447. http://www.jstor.org/stable/20125810.
Silver, D. (2015). Lecture 1: Introduction to reinforcement learning. *Google DeepMind*, *1*, 1–10.
Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., & Hassabis, D. (2016). Mastering the game of go with deep neural networks and tree search. *nature*, *529*(7587), 484–489.
Sitzmann, V., Martel, J., Bergman, A., Lindell, D., & Wetzstein, G. (2020). Implicit neural representations with periodic activation functions. *Advances in Neural Information Processing Systems*, *33*, 7462–7473.
Sloman, A. (2019). The computer revolution in philosophy: Philosophy, science and models of mind.
Taigman, Y., Yang, M., Ranzato, M., & Wolf, L. (2014). DeepFace: Closing the gap to human-level performance in face verification. *Proceedings of sthe IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. https://doi.org/10.1109/CVPR.2014.220
Van Es, T. (2021). Living models or life modelled? On the use of models in the free energy principle. *Adaptive Behavior*, *29*(3), 315–329.
Van Gelder, T. (1995). What might cognition be, if not computation? *The Journal of Philosophy*, *92*(7), 345–381.
Varela, F. J. (1996). Neurophenomenology: A methodological remedy for the hard problem. *Journal of Consciousness Studies*, *3*(4).
Varela, F. J., Thompson, E., & Rosch, E. (1991). *The embodied mind: Cognitive science and human experience*. MIT Press.
Von der Malsburg, C. (1995). Binding in models of perception and brain function. *Current opinion in neurobiology*, *5*(4), 520–526.
Wang, X., Wang, S., Liang, X., Zhao, D., Huang, J., Xu, X., & Miao, Q. (2022). *Deep reinforcement learning: A survey*. IEEE Transactions on Neural Networks and Learning Systems.
Xu, J., de Mello, S., Liu, S., Byeon, W., Breuel, T., Kautz, J., Wang, X., & San Diego, U. (2022). GroupViT: Semantic Segmentation Emerges from Text Supervision. In *openaccess.thecvf.com*. https://github.com/NVlabs/GroupViT.
Yamins, D. L. K., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience (Vol*, *19*(3), https://doi.org/10.1038/nn.4244.

Yoshimi, J. (2011). Phenomenology and connectionism. *Frontiers in psychology*, *2*, 288.
Zahavi, D. (2008). Phenomenology. *The Routledge companion to twentieth century philosophy* (pp. 661–692). Routledge.
Zahavi, D. (2018). Brain, mind, World: Predictive Coding, Neo-Kantianism, and Transcendental Idealism. *Husserl Studies*, *34*(1), https://doi.org/10.1007/s10743-017-9218-z.
Zhang, Y., Tino, P., Leonardis, A., & Tang, K. (2021). A survey on neural network interpretability. In *IEEE Transactions on Emerging Topics in Computational Intelligence* (Vol. 5, Issue 5). https://doi.org/10.1109/TETCI.2021.3100641