frontiers in Computational Neuroscience

HYPOTHESIS AND THEORY
published: 22 December 2020
doi: 10.3389/fncom.2020.556544

Will We Ever Have Conscious Machines?

Patrick Krauss[1,2]* *and Andreas Maier*[3]

[1] Neuroscience Lab, University Hospital Erlangen, Erlangen, Germany, [2] Cognitive Computational Neuroscience Group, Chair of Linguistics, Friedrich-Alexander University Erlangen-Nürnberg (FAU), Erlangen, Germany, [3] Chair of Machine Intelligence, Friedrich-Alexander University Erlangen-Nürnberg (FAU), Erlangen, Germany

The question of whether artificial beings or machines could become self-aware or conscious has been a philosophical question for centuries. The main problem is that self-awareness cannot be observed from an outside perspective and the distinction of being really self-aware or merely a clever imitation cannot be answered without access to knowledge about the mechanism's inner workings. We investigate common machine learning approaches with respect to their potential ability to become self-aware. We realize that many important algorithmic steps toward machines with a core consciousness have already been taken.

Keywords: machine consciousness, artificial intelligence, theories of consciousness, deep learning, machine learning, philosophy of mind, global workspace, correlates of consciousness

1. INTRODUCTION

The question of understanding consciousness is in the focus of philosophers and researchers for more than two millennia. Insights range broadly from *"Ignorabimus"*—*"We will never know*[1]*."* to mechanistic ideas with the aim to construct artificial consciousness following Richard Feynman's famous words *"What I cannot create, I do not understand*[2]*."*

The major issue that precludes the analysis of consciousness is its subjectivity. Our mind is able to feel and process our own conscious states. By induction, we are also able to ascribe conscious processing to other human beings. However, once we try to imagine to be another species, as Nagel describes in his seminal work "What is it like to be a Bat?"(Nagel, 1974), we immediately fail to follow such experience consciously.

Another significant issue is that we are not able to determine consciousness by means of behavioral observations as Searle demonstrates in his thought experiment (Searle, 1980). Searle describes a room that we cannot enter. One can pass messages written in Chinese to the room and the room returns messages to the outside world. All the messages and questions passed to the room are answered correctly as a Chinese person would. A first conclusion would be that there is somebody in the "Chinese Room" who speaks Chinese and answers the questions. However, the person in the room could also have simply access to a large dictionary that contains all possible questions and the respective answers. When we are not able to understand how the information is actually processed, we will never be able to determine whether a system is conscious or not. In this article, we want to explore these and different thoughts in literature in order to address the problem of consciousness.

[1] With this simple statement, Emil du Bois-Reymond concluded his talk on the limits of scientific knowledge about the relation of brain processes and subjective experience at the 45th annual meeting of German naturalists and physicians in 1872.
[2] Richard Feynman left these words on his blackboard in 1988 at the time of his death as a final message to the world.

OPEN ACCESS

Edited by:
Si Wu,
Peking University, China

Reviewed by:
Valeri Makarov,
Complutense University of Madrid, Spain
Maurizio Mattia,
National Institute of Health (ISS), Italy

***Correspondence:**
Patrick Krauss
patrick.krauss@uk-erlangen.de

Received: *06 May 2020*
Accepted: *26 November 2020*
Published: *22 December 2020*

Citation:
Krauss P and Maier A (2020) Will We Ever Have Conscious Machines? Front. Comput. Neurosci. 14:556544. doi: 10.3389/fncom.2020.556544

Since there exist three different, largely isolated groups in the scientific community aiming to investigate consciousness, i.e., philosophy, neuroscience, and computer science, here, we try to overcome the mutual gaps between these complementary camps of research, by incorporating arguments from each side, thereby providing a balanced overview of consciousness research. We therefore revisit works in philosophy, neuroscience, artificial intelligence, and machine learning. Following the new paradigm of *cognitive computational neuroscience* (Kriegeskorte and Douglas, 2018), we present how the convergence of these fields could potentially also lead to new insights regarding consciousness.

Since philosophy is the root of all scientific research, and in particular the research on consciousness, it was frequently argued that there is a need for more philosophical thinking in scientific research in order to be able to ask the right questions (Thagard, 2009; Rosen, 2015; Laplane et al., 2019). Therefore, we start with the philosophical perspective, and put a special emphasis on the description of the most important arguments and positions from the philosophy of mind.

2. THE PHILOSOPHICAL PERSPECTIVE

More than two thousand years ago, Aristotle was convinced that only humans are endowed with a rational soul. All animals, however, live only with the instincts necessary for survival, like biological automata. Along the same line, in the statement "*Cogito ergo sum*" also Descartes realized being self-aware is reserved for human beings. In his view, this insight is fundamental for any philosophical approach (Descartes, 1990).

Modern philosophy went on to differentiate the problem into an easy and a hard problem. While the "easy problem" is to explain its function, dynamics, and structure, the "hard problem of consciousness" Chalmers (1995) is summarized in the Internet Encyclopedia of Philosophy (Weisberg, 2020) as:

> "The hard problem of consciousness is the problem of explaining why any physical state is conscious rather than nonconscious. It is the problem of explaining why there is "something it is like" for a subject in conscious experience, why conscious mental states "light up" and directly appear to the subject."

In order to avoid confusion some scientists prefer to speak of "conscious experience" or only "experience" instead of consciousness (Chalmers, 1995). As already noted, the key problem of deriving models of conscious events is that they can only be perceived subjectively. As such it is difficult to encode such an experience in a way that it can be recreated by others. This gives rise to the so-called "qualia problem" (Crane, 2012) as we can never be sure, e.g., that the color red consciously looks the same to another person. Extension of this line of thought leads again to Nagel's thought experiment (Nagel, 1974).

According to (Weisberg, 2020), approaches to tackle the problem from a philosophical point of view are very numerous, but none of them can be considered to be exhaustive:

- **Eliminativism** (Rey, 1988) demonstrates that the mind is fully functional without the experience of consciousness. Being non-functional, consciousness can be neglected.
- The view of **strong reductionism** proposes that consciousness can be deconstructed into simpler parts and be explained by functional processes. Such considerations gave rise to the Global Work Space Theory (Newman and Baars, 1993; Baars, 1994; Baars and Newman, 1994) or Integrated Information Theory (Tononi, 2004, 2008) in neuroscience. The main critique of this view, is that any mechanistic solution to consciousness that is not fully understood will only mimic true consciousness, i.e., one could construct something that appears conscious that simply isn't as the Chinese Room argument demonstrates (Searle, 1980).
- **Mysterianism** proposes that the question of consciousness cannot be tackled with scientific methods. Therefore any investigation is in vain and the explanatory gap cannot be closed (Levine, 2001).
- In **Dualism** the problem is tackled as consciousness being metaphysical that is independent of physical substance (Descartes, 1990). Modern versions of Dualism exist, but virtually all of them require to reject that our world can be fully described by physical principles. Recently, Penrose and Hammeroff tried to close this gap using quantum theory (Penrose, 1994; Hameroff and Penrose, 2014). We dedicate a closer description of this view in a later section of this article.
- Assuming that metaphysical world and physical world simply do not interact does not require to reject physics and gives rise to **Epiphenomenalism** (Campbell, 1992).

There are further theories and approaches to address the hard problem of consciousness that we do not want to detail here. To the interested reader, we recommend to study the Internet Encyclopedia of Philosophy (Weisberg, 2020) as further reading into the topic.

In conclusion, we observe that a major disadvantage of exploring the subject of consciousness by philosophical means is that we will never be able to explore the inside of the Chinese Room. Thought alone will not be able to open the black box. Neuroscience, however, offers various approaches to explore the inside by means of measurement, which might be suitable to tackle the problem.

3. CONSCIOUSNESS IN NEUROSCIENCE

In 1924, Hans Berger recorded, for the first time, electrical brain activity using electroencephalography (EEG) (Berger, 1934). This breakthrough enabled the investigation of different mental states by means of electrophysiology, e.g., during perception (Krauss et al., 2018a) or during sleep (Krauss et al., 2018b). The theory of cell assemblies, proposed by Hebb (1949), marked the starting point for the scientific investigation of neural networks as the biological basis for perception, cognition, memory, and action. In 1965, Gazzaniga demonstrated that dissecting the corpus callosum which connects the two brain hemispheres with each other results in a split of consciousness (Gazzaniga et al., 1965; Gazzaniga, 2005). Almost ten years later, Weiskrantz et al.

discovered a phenomenon for which the term "blindsight" has been coined: following lesions in the occipital cortex, humans loose the ability to consciously perceive, but are still able to react to visual stimuli (Weiskrantz et al., 1974; Weiskrantz and Warrington, 1975). In 1983, Libet demonstrated that voluntary acts are preceded by electrophysiological readiness potentials that have their maximum at about 550 *ms* before the voluntary behavior (Libet et al., 1983). He concluded that the role of conscious processing might not be to initiate a specific voluntary act but rather to select and control volitional outcome (Libet, 1985). In contrast to the above mentioned philosophical tradition from Aristotle to Descartes that consciousness is a phenomenon that is exclusively reserved for humans, in contemporary neuroscience most researchers tend to regard consciousness as a gradual phenomenon, which in principle also occurs in animals (Boly et al., 2013), and several main theories of how consciousness emerges have been proposed so far.

3.1. Neural Correlates of Consciousness

Based on Singer's observation that high-frequency oscillatory responses in the feline visual cortex exhibit inter-columnar and inter-hemispheric synchronization which reflects global stimulus properties (Gray et al., 1989; Engel et al., 1991; Singer, 1993) and might therefore be the solution for the so called "binding problem" (Singer and Gray, 1995), Crick and Koch suggested Gamma frequency oscillations to play a key role in the emergence of consciousness (Crick and Koch, 1990). Koch further developed this idea and investigated neural correlates of consciousness in humans (Tononi and Koch, 2008; Koch et al., 2016). He argued that activity in the primary visual cortex, for instance, is necessary but not sufficient for conscious perception, since activity in areas of extrastriate visual cortex correlates more closely with visual perception, and damage to these areas can selectively impair the ability to perceive particular features of stimuli (Rees et al., 2002). Furthermore, he discussed the possibility that the timing or synchronization of neural activity might correlate with awareness, rather than simply the overall level of spiking (Rees et al., 2002). A finding which is supported by recent neuroimaging studies of visual evoked activity in parietal and prefrontal cortex areas (Boly et al., 2017). Based on these findings, Koch and Crick provided a framework for consciousness, where they proposed a coherent scheme to explain the neural activation of visual consciousness as competing cellular clusters (Crick and Koch, 2003). Finally, the concept of neural correlates of consciousness has been further extended to an index of consciousness based on brain complexity (Casarotto et al., 2016), which is independent of sensory processing and behavior (Casali et al., 2013), and might be used to quantify consciousness in comatose patients (Seth et al., 2008). While such approaches, known as *perturbational complexity index* (Casali et al., 2013), are designed to assess dynamical or processing complexity, they are not adequate to measure the underlying connectivity or circuitry.

3.2. Consciousness as a Computational Phenomenon

Motivated by the aforementioned findings concerning the neural correlates of consciousness, Tononi introduced the concept of integrated information, which according to his "Integrated Information Theory of Consciousness" plays a key role in the emergence of consciousness (Tononi, 2004, 2008). This theory represents one of two major theories of contemporary research in consciousness. According to this theory, the quality or content of consciousness is identical to the form of the conceptual structure specified by the physical substrates of consciousness, and the quantity or level of consciousness corresponds to its irreducibility, which is defined as integrated information (Tononi et al., 2016).

Tegmark generalized Tononi's framework even further from neural-network-based consciousness to arbitrary quantum systems. He proposed that consciousness can be understood as a state of matter with distinctive information processing abilities, which he calls "perceptronium," and investigates interesting links to error-correcting codes and condensed matter criticality (Tegmark, 2014, 2015).

Even though, there is large consensus that consciousness can be understood as a computational phenomenon (Cleeremans, 2005; Seth, 2009; Reggia et al., 2016; Grossberg, 2017), there is dissent about which is the appropriate level of granularity of description and modeling (Kriegeskorte and Douglas, 2018). Penrose and Hameroff even proposed that certain features of quantum coherence could explain enigmatic aspects of consciousness, and that consciousness emerges from brain activities linked to fundamental ripples in spacetime geometry. In particular, according to their model of orchestrated objective reduction (Orch OR), they hypothesize that the brain is a kind of quantum computer, performing quantum computations in the microtubeles, which are cylindrical protein lattices of the neurons' cytoskeleton (Penrose, 1994; Hameroff and Penrose, 1996; Hameroff, 2001).

However, Tegmark and Koch argue, that the brain can be understood within a purely neurobiological framework, without invoking any quantum-mechanical properties: quantum computations which seek to exploit the parallelism inherent in entanglement, require that the qubits are well-isolated from the rest of the system, whereas on the other hand, coupling the system to the external world is necessary for the input, the control, and the output of the computations. Due to the wet and warm nature of the brain, all these operations introduce noise into the computation, which causes decoherence of the quantum states, and thus makes quantum computations impossible. Furthermore, they argue that the molecular machines of the nervous system, such as the pre- and post-synaptic receptors, are so large that they can be treated as classical rather than quantum systems, i.e., that there is nothing fundamentally wrong with the current classical approach to neural network simulations (Tegmark, 2000; Koch and Hepp, 2006, 2007).

3.3. The Global Workspace Theory

In the 1990s, Baars introduced the concept of a virtual "Global Workspace" that emerges by connecting different brain areas (**Figure 1**) to describe consciousness (Newman and Baars, 1993; Baars, 1994, 2007; Baars and Newman, 1994). This idea was taken up and further developed by Dehaene (Dehaene et al., 1998, 2011, 2014; Dehaene and Naccache, 2001; Dehaene and

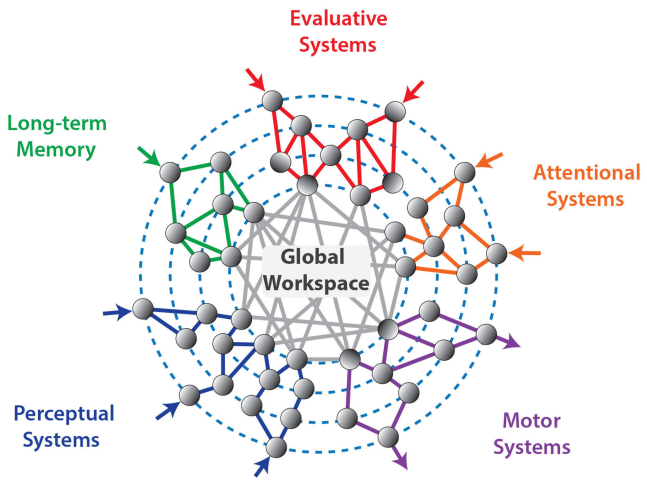

FIGURE 1 | The Global Workspace emerges by connecting different brain areas according to Dehaene.

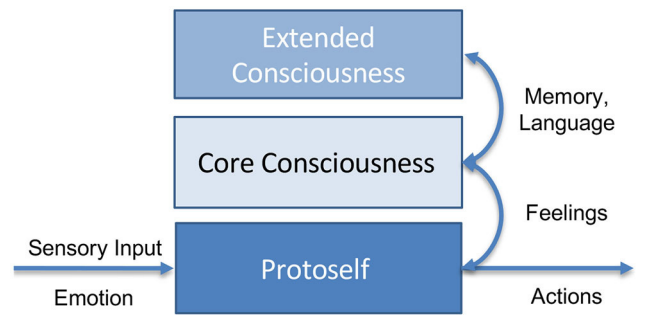

FIGURE 2 | Simplified view of Damasio's model of consciousness: The protoself processes emotions and sensory input unconsciously. Core consciousness arises from the protoself which allows to put the itself into relation. Projections of emotions give rise to higher-order feelings. With access to memory and extended functions such as language processing the extended consciousness emerges.

Changeux, 2004; Sergent and Dehaene, 2004). Today, besides the Integrated Information Theory, the Global Workspace Theory represents the second major theory of consciousness, being intensively discussed in the field of cognitive neuroscience. Based on the implications of this theory, i.e., that *"consciousness arises from specific types of information-processing computations, which are physically realized by the hardware of the brain"* (Dehaene et al., 2017), Dehaene argues that a machine endowed with these processing abilities *"would behave as though it were conscious; for instance, it would know that it is seeing something, would express confidence in it, would report it to others, could suffer hallucinations when its monitoring mechanisms break down, and may even experience the same perceptual illusions as humans"* (Dehaene et al., 2017). Indeed, it has been demonstrated recently that artificial neural networks trained on image processing can be subject to the same visual illusions as humans (Gomez-Villa et al., 2018; Watanabe et al., 2018; Benjamin et al., 2019)

3.4. Damasio's Model of Consciousness

Damasio's model of consciousness was initially published in his popular science book "The feeling of what happens" (Damasio, 1999). Later Damasio also published the central ideas in peer-reviewed scientific literature (Damasio and Meyer, 2009). With the ideas being published first in a popular science book, most publications on consciousness neglect his contributions. However, we believe that his thoughts deserve more attention. Therefore, we want to introduce his ideas quickly in this section.

The main idea in Damasio's model is to relate consciousness to the ability to identify one's self in the world and to be able to put the self in relation with the world. However, a formal definition is more complex and requires the introduction of several concepts first.

He introduces three levels of conscious processing:

- The fundamental **protoself** does not possess the ability to recognize itself. It is a mere processing chain that reacts to inputs and stimuli like an automaton, completely non-conscious. As such any animal has a protoself according to this definition. However, also more advanced lifeforms including humans exhibit this kind of self.
- A second stage of consciousness is the **core consciousness**. It is able to anticipate reactions in its environment and adapts to them. Furthermore, it is able to recognize itself and its parts in its own image of the world. This enables it to anticipate and to react to the world. However, core consciousness is also volatile and not able to persist for hours to form complex plans.
  In contrast to many philosophical approaches, core consciousness does not require to represent representations of the world in words or language. In fact, Damasio believes that progress in understanding conscious processing has been impeded by dependence on words and language.
- The **extended consciousness** enables human-like interaction with the world. It builds on top of core consciousness and enables further functions such as access to memory in order to create an autobiographic self. Also being able to process words and language falls into the category extended consciousness and can be interpreted as a form of serialization of conscious images and states.

In Damasio's theory emotions and feelings are fundamental concepts (Damasio, 2001). In particular Damasio differentiates emotions from feelings. **Emotions** are direct signals that indicate a positive or negative state of the (proto-)self. **Feelings** emerge only in conjunction with images of the world and can be interpreted as a second-order emotion that is derived from the world representation and future possible events in the world. Both are crucial for the emergence of consciousness. **Figure 2** schematically puts the described terms in relation.

After having defined the above concepts, Damasio now goes on to attempt and describe a model of (core) consciousness. In his theory, consciousness does not merely emerge from the ability to identify oneself in the world or an image of the world. For conscious processing, additionally feeling oneself in the sense of desiring to exist is required. Hence, he postulates a feeling, i.e., a derived second-order emotion, between the protoself and its internal representation of the world. Conscious beings as

such want to identify oneself in the world and want to exist. From an evolutionary perspective as he argues, this is possibly a mechanism to enforce self-preservation.

In the context of this article, Damasio's theory is interesting for two major reasons. On the one hand, it describes a biologically plausible model of consciousness, as he assigns all stages of consciousness to certain structures in the brain, and associates them to the respective function. On the other hand, Damasio's model is mechanistic, thus it can be, at least in principle, completely implemented as a computer program.

Summing up, we can conclude that neuroscience is able to describe fundamental processes in the brain that give rise to complex phenomena such as consciousness. However, the different methods of observation in neuroscience are still not sufficient. Neither EEG nor fMRI nor any other contemporary imaging method provide a temporal and spatial resolution that is even close to be enough fine-grained to observe what exactly is happening in the brain *in-vivo* (Maier et al., 2018), i.e., the human brain consisting of about 8.6×10^{10} neurons (Herculano-Houzel, 2009) interconnected by approximately 10^{15} synapses (Sporns et al., 2005; Hagmann et al., 2008) is far away from being entirely accessible. At this point, the recent massive progress in artificial intelligence and machine learning, especially deep learning, comes to our attention. In contrast to the human brain, artificial neural networks or any other computational model provide the decisive advantage of being fully accessible at any time, i.e., the state of each parameter can be read out without any restrictions with respect to precision. Furthermore, there is increasing evidence that, even though most artificial neural networks largely lack biological plausibility, they are nevertheless well-suited for modeling brain function. A number of recent studies have shown striking similarities in the processing and representational dynamics between artificial neural networks and the brain (Cichy et al., 2016; Zeman et al., 2020). For instance, in deep neural networks trained on visual object recognition, the spontaneous emergence of number detectors (Nasr et al., 2019), solid shape coding (Srinath et al., 2020), or center-periphery spatial organization (Mohsenzadeh et al., 2020) was observed. Furthermore, grid-like representations known to exist in the entorhinal cortex (Hafting et al., 2005) spontaneously emerge in recurrent neural networks trained to perform spatial localization (Cueva and Wei, 2018) or navigation tasks (Banino et al., 2018).

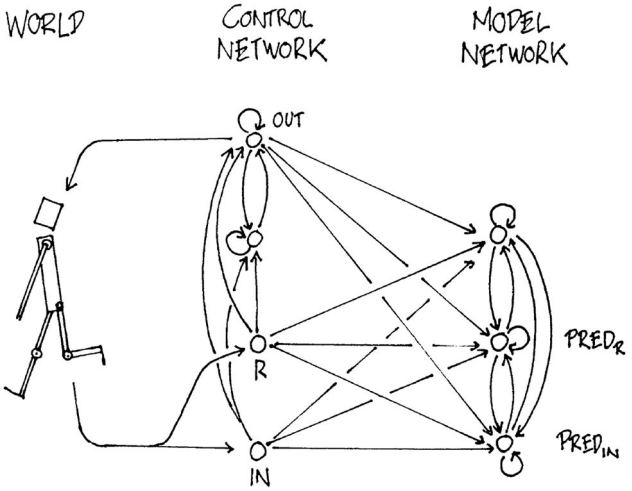

FIGURE 3 | Schmidhuber already proposed a first model for autonomous agents in 1990 (Schmidhuber, 1990). Similar to ideas presented by Damasio, the model receives a reward *R* and input *IN* from the world. The network processes the input, does predictions on the world $PRED_{IN}$ and predictions about future rewards $PRED_R$. Finally, actions are undertaken in *OUT*. Reprinted with permission.

4. CONSCIOUSNESS IN ARTIFICIAL INTELLIGENCE

In artificial intelligence (AI) numerous theories of consciousness exist (Sun and Franklin, 2007; Starzyk and Prasad, 2010). Implementations often focus on the Global Work Space Theory with only limited learning capabilities (Franklin and Graesser, 1999), i.e., most of the consciousness is hard-coded and not trainable (Kotov, 2017). An exception is the theory by van Hateren which closely relates consciousness to simultaneous forward and backward processing in the brain (van Hateren, 2019). Yet, algorithms that were investigated so far made use of a global work space and mechanistic hard-coded models of consciousness. Following this line, research on minds and consciousness rather focuses on representation than on actual self-awareness (Tenenbaum et al., 2011). Although representation will be important to create human-like minds and general intelligence (Gershman et al., 2015; Lake et al., 2017; Mao et al., 2019), a key factor to become conscious is the ability to identify a *self* in one's environment (Dehaene et al., 2017). A major drawback of pure mechanistic methods, however, is that the complete knowledge on the model of consciousness is required in order to realize and implement them. As such, in order to develop these models to higher forms such as Damasio's extended consciousness, a complete mechanistic model of the entire brain including all connections is required.

4.1. Consciousness in Machine Learning

A possible solution to this problem is machine learning, as it allows to form and train complex models. The topic of consciousness, however, is neglected in the field to a large extent. On the one hand, this is because of the concerns that the brain and consciousness will never be successfully simulated in a computer system (Penrose, 2001; Hameroff and Penrose, 2014). On the other hand, consciousness is considered to be an extremely hard problem and current results in AI are still meager (Brunette et al., 2009).

The state-of-the-art in machine learning instead focuses on supervised and unsupervised learning techniques (Bishop, 2006). Another important research direction is reinforcement learning (Sutton and Barto, 2018) that aims at learning of suitable actions for an agent in a given environment. As consciousness is

often regarded to be associated with embodiment, reinforcement learning is likely to be important for modeling of consciousness.

The earliest work that the authors are aware of attempting to model and create agents that learn their own representation of the world entirely using machine learning date back to the early 1990's. Already in 1990, Schmidhuber proposed a model for dynamic reinforcement learning in reactive environments (Schmidhuber, 1990) and found evidence for self-awareness in 1991 (Schmidhuber, 1991). The model follows the idea of a global work space. In particular, future rewards and inputs are predicted using a world model as shown in **Figure 3**. Yet, Schmidhuber was missing a theory on how to analyse intelligence and consciousness in this approach. Similar to Tononi (2008), Schmidhuber followed the idea of compressed neural representation. Interestingly, compression is also key to inductive reasoning, i.e., learning from few examples which we typically deem as intelligent behavior.

Solomonoff's *Universal Theory of Inductive Inference* (Solomonoff, 1964) gives a theoretic framework to inductive reasoning. It combines information theory with compression theory and results in a formalization of Occam's razor preferring simple models over complex ones (Maguire et al., 2016), as simple models are more likely from an information theoretic point of view[3].

Under Schmidhuber's supervision, Hutter applied Solomonoff's theory to machine learning to form a theory of *Universal Artificial Intelligence* (Hutter, 2004). In this theory, intelligent behavior stems from *efficient compression*[4] of inputs, e.g., from the environment, such that predictions and actions are performed optimally. Again, models capable of describing a global work space play an important role.

Maguire et al. further expand on this concept to extend Solomonoff's and Hutter's theories to also describe consciousness. Following the ideas of Tononi and Koch (Rees et al., 2002) consciousness is understood as data compression, i.e., the optimal integration of information (Maguire et al., 2016). The actual consciousness emerges from binding of information and is inherently complex. As such, consciousness can also not be deconstructed into mechanical sub-components, as the decomposition would destroy the sophisticated data compression. Maguire et al. even provide a mathematical proof to demonstrate that consciousness is either integrated and therefore cannot be decomposed or there is an explicit mechanistic way of modeling and describing consciousness (Maguire et al., 2016).

Based on the extreme success of deep learning (LeCun et al., 2015), also several scientists observed similarities in neuroscience and machine learning. In particular, deep learning allows to build complex models that are hard to analyse and interpret at the benefit of making complex predictions. As such both fields are likely to benefit each other in the ability to understand and interpret complex dynamic systems (Marblestone et al., 2016; Hassabis et al., 2017; Van Gerven, 2017; Kriegeskorte and Douglas, 2018; Barrett et al., 2019; Richards et al., 2019; Savage, 2019). In particular, hard-wiring following biological ideas might help to reduce the search space dramatically (Zador, 2019). This is in line with recent theoretical considerations in machine learning as prior knowledge allows to reduce maximal error bounds (Maier et al., 2019b). Both fields can benefit from these ideas as recent discoveries of e.g., successor representation show (Stachenfeld et al., 2017; Gershman, 2018; Geerts et al., 2019). Several scientists believe that extension of this approach to social, cultural, economic, and political sciences will create even more synergy resulting in the field of *machine behavior* (Rahwan et al., 2019).

5. CAN CONSCIOUSNESS EMERGE IN MACHINE LEARNING SYSTEMS?

After having reviewed philosophy, neuroscience, and the state-of-the-art in AI and machine learning, we can now analyse the most important concepts in the field of machine learning, especially deep learning, to assess whether they have the potential to create consciousness following one of the previous theories. In particular, we focus on the ability of the system to represent a category of self and how this self-awareness is constructed, as all theories of consciousness require at least experiencing the self.

In **Figure 4**, we provide an overview of important models depicted as box-and-arrow schemes, following the standard way to communicate neural network architectures within the machine learning community. We denote *feed-forward connections*[5] as dashed black lines, *recurrent connections*[6] as solid

[3] Note that of course, conscious brains have to be complex in order to implement a computing mechanism that is able to realize the process of model creation and selection accordingly. Thus, the concept of Occam's razor is by no means in contrast to this conviction. Rather, it describes the selection procedure of the models itself. According to Occam's razor, simple models should be preferred over more complex ones. The idea behind this approach is that a good model should, on the one hand be as complex as necessary, but on the other hand as simple as possible, i.e., not too complex, e.g., containing too many free parameters, that are not required to capture the phenomenon under consideration.

[4] In computer science, the concept of compression, or compressibility respectively, does not imply simplicity of the input or the model architectures. On the contrary, deriving more and more compressed representations from the input, requires highly complex processes for which in turn complex architectures are crucial. For instance, a random sequence of digits is incompressible but not complex. On the other hand, the infinitely many digits of π can be compressed to algorithms, i.e., representations, of finite length that (re-)produce the digit sequence of π with arbitrary precision. Thus, such an algorithm of finite length is both an extremely compressed representation of π, and highly complex.

[5] Depending on the specific task and input data, these connections are adjusted according to a certain learning algorithm (Bishop, 2006), e.g., supervised with error backpropagation (Chauvin and Rumelhart, 1995), unsupervised with, e.g., contrastive divergence (Carreira-Perpinan and Hinton, 2005) or through reinforcement learning (Sutton and Barto, 2018).

[6] These connections are hard-wired, i.e., not adjusted during training, and provide a copy of the current module's internal state to the same module. Thus, in each subsequent processing step, the module receives the current input together with its own previous internal state. This design principle enables processing of temporal structured data. So called recurrent neural networks (Tsoi and Back, 1997), like e.g., long short-term memories (LSTMs) (Hochreiter and Schmidhuber, 1997) have proven to do extremely well in speech and language processing, for instance.

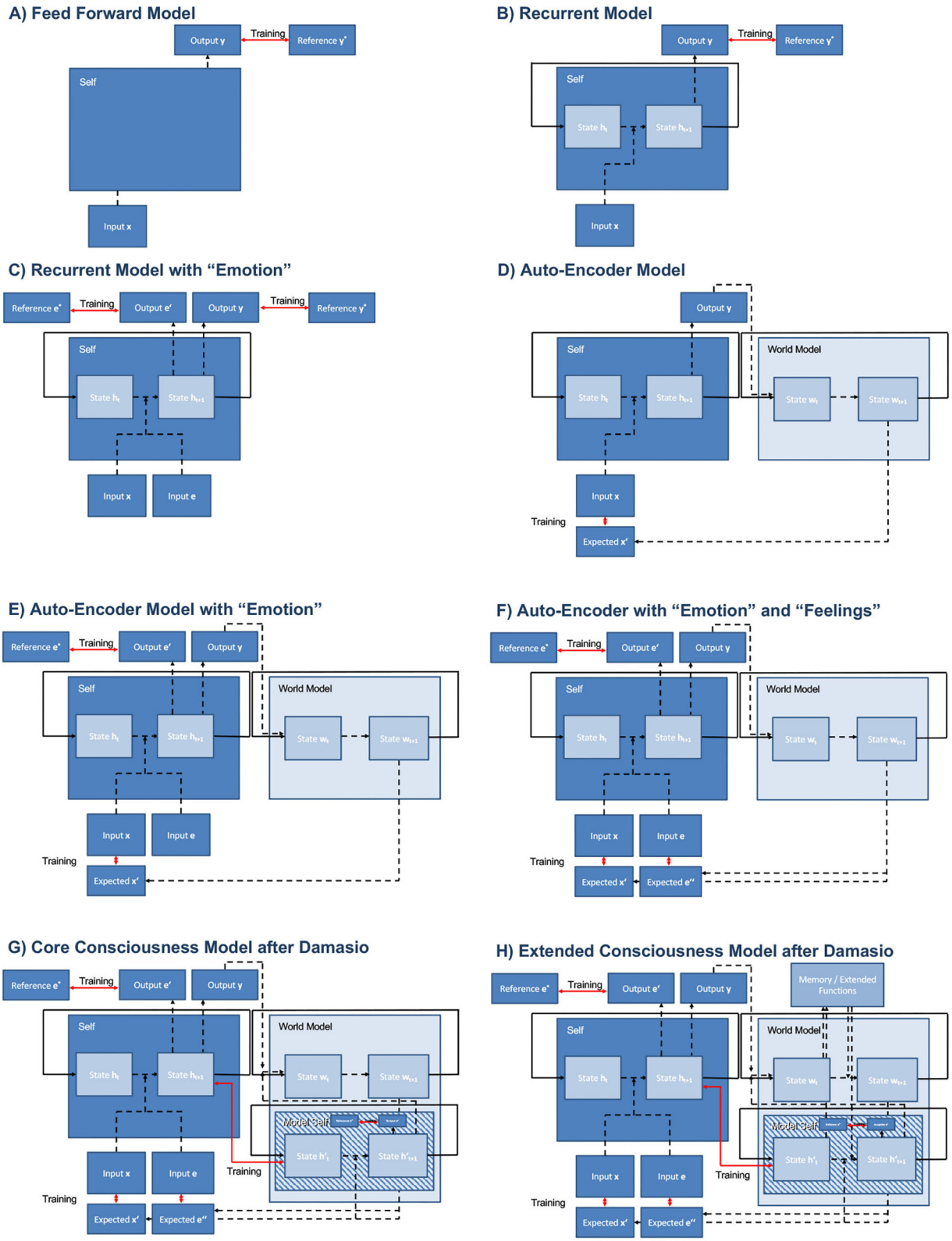

FIGURE 4 | Overview on typical architectures in machine learning and proposed models. Hard-coded paths, i.e., recurrent connections, are indicated by solid black lines, trainable feed-forward connections by dashed black lines, and training losses by red lines. Arrows indicate information flow. While architectures **(A–E)** do not match theories of consciousness, architectures **(F–H)** implement theories by Schmidhuber and Damasio.

black lines, and *training losses*[7] as red lines. Arrows indicate the direction of information flow.

The trainable feed-forward connections perform certain transformations on the input data yielding the output. In the most simple case, each depicted feed-forward connection could either be realized as a direct link with trainable weights from the source module to the respective target module, a so called perceptron (Minsky and Papert, 2017), or with a single so called hidden layer between source and target module. Those multi-layer neural networks are known to be universal function approximators (Hornik, 1991). Without loss of generality, the depicted feed-forward connections could also be implemented by other deep feed-forward architectures (Maier et al., 2019a) comprising several stacked hidden layers, and could thus be inherently complex[8].

Figure 4A shows a simple feed-forward architecture which can be trained through supervised learning, i.e., it requires external labeled training data $\mathbf{y}^*$ to adjust its trainable weights given input $\mathbf{x}$ to produce output $\mathbf{y}$. During training the error between expected output $\mathbf{y}^*$ and actual output $\mathbf{y}$ is minimized. Models like this are used in machine learning to classify static input like images for instance (LeCun et al., 2015). In this model, we do not expect the emergence of consciousness.

Figure 4B shows a similar setup, yet with an additional recurrent connect feeding the momentary internal state $\mathbf{h_t}$ back to the input. Thus, the subsequent internal state $\mathbf{h_{t+1}}$ and output $\mathbf{y_{t+1}}$ depend not only on the subsequent input $\mathbf{x_{t+1}}$ but also on the previous internal state. Note that we only depict a simple recurrent cell here with time-dependent internal state $\mathbf{h_t}$. Without loss of generality, this could also be realized by more complex architectures like gated recurrent units (GRUs) (Cho et al., 2014) or long short-term memory cells (LSTMs) (Hochreiter and Schmidhuber, 1997). Again, models like this fall into the category of supervised learning, i.e., require labeled input data or information about the expected output, respectively. Due to their ability of processing sequential input, these models are widely used in contemporary machine learning, e.g., natural language processing (Young et al., 2018). The emergence of consciousness in those models is neither observed, nor supported by any of the theories presented so far.

In **Figure 4C**, we introduce the concept of "Emotion" following Damasio's wording. In machine learning terms, this reflects an additional loss. Now, the system receives an additional input $\mathbf{e}$ that is associated to a valence or value, i.e., a reward. Without loss of generality, we can assume positive entries in $\mathbf{e}$ to be associated to desirable states for the system and negative values to undesirable states. As such, training using $\mathbf{e}$ falls into the category of reinforcement learning (Sutton and Barto, 2018) that aims at maximizing future rewards of $\mathbf{e}$, by adjusting or choosing appropriate output $\mathbf{y}$, i.e., action or behavior. In order to model competing interests and saturation effects, e.g., a full battery does not need to be charged further, we introduce a reference $\mathbf{e}^*$, i.e., a desired reward, that is able to model such effects. This corresponds to the concept of homeostasis in biology: desired rewards $\mathbf{e}^*$ correspond to preferred values of metabolic parameters like blood oxygenation level for instance, whereas the actual rewards $\mathbf{e}$ would correspond to actual values of such metabolic parameters. Like in a feedback loop, the organism seeks to minimize the difference between desired and actual values of metabolic parameters. Note that we deem the system to be able to predict the expected future reward $\mathbf{e}'$ from its current state $\mathbf{h_t}$ following a so called deep Q-learning paradigm (Mnih et al., 2015). Here we use $\mathbf{e}'$ and $\mathbf{e}^*$ to construct a trainable reinforcement loss, to be able to learn from low-level rewards $\mathbf{e}$ to adjust the organism's output action or behavior, respectively. In Damasio's model, the totality of all actual low-level rewards $\mathbf{e}$ correspond to emotions. Although being able to learn, systems like this still need supervision to train the weights producing appropriate output (action / behavior) $\mathbf{y}$ using a reference. In machine learning, such models fall into the class of model-free reinforcement learning (Sutton and Barto, 2018). Although these models are able to learn playing computer games (Mnih et al., 2015) or board games like Go (Silver et al., 2017) at human level, the emergence of consciousness is not expected, let alone observed, since this setup also does not match any theory of consciousness so far.

As self-awareness is a requirement for base consciousness, we deem a world model, i.e., a model of the organism's environment, to be necessary. Such an approach is shown in **Figure 4D**, and corresponds to so called auto-encoders which are used in machine learning for dimensionality reduction (Wang et al., 2016) or the construction of compact feature spaces, so called embeddings (Lange and Riedmiller, 2010). Given the organism's produced output (action / behavior) $\mathbf{y}$, the world model is used to estimate an expected future input $\mathbf{x}'$. In the figure, we chose a recurrent model capturing the sequence of external states of the world $\mathbf{w_t}$ that is independent of the sequence of internal states of the agent $\mathbf{h_t}$. To gain consciousness, this model misses at least a link from internal to external state and "emotions" that would guide future decisions.

Combining the two previously described models (**Figures 4C,D**), i.e., adding low-level rewards/emotions to the auto-encoder model results in the model shown in **Figure 4E**, which corresponds, in contrast to model-free reinforcement learning, to model-based reinforcement learning (Sutton and Barto, 2018). Again world and self are disconnected, hence inhibiting self-representation and self-discovery. Approaches like this are already being explored for video game control (Kaiser et al., 2019).

[7]Training losses provide a reference during learning to adjust weights in trainable connections. In supervised learning (Bishop, 2006), the difference between actual and expected output, i.e., the error, serves as training loss (Chauvin and Rumelhart, 1995). In contrast, in reinforcement learning (Sutton and Barto, 2018), rewards indicating the appropriateness of a certain action or output are provided as feed-back.

[8]In principle, all presented models could be implemented, using state-of-the-art machine learning libraries such as Keras (Chollet, 2015), Tensorflow (Abadi et al., 2016), or PyTorch (Paszke et al., 2019). Of course, a lot of experiments are still necessary and some of the presented models are still difficult to train. However, some work toward this direction has already been done. For instance, some neural network based AI models capable of playing classic Atari games at human level do already incorporate world models of the games' environments (Kaiser et al., 2019), i.e., make use of so called model-based reinforcement learning (Sutton and Barto, 2018).

With a world-model being present, besides predicting future rewards $\mathbf{e}'$ from the internal agent state, we are now able to additionally predict future rewards $\mathbf{e}''$ that also take into account the external state of the world and the chosen action / behavior **y**. As such **Figure 4F** is the first one that would implement a trainable version of deep Q-learning. However, development of consciousness is debatable, as the model does not feature a link between the external state of the world $\mathbf{w_t}$ and the internal state of the agent $\mathbf{h_t}$. If we would add trainable connections from $\mathbf{h_t}$ to $\mathbf{w_t}$ and vice versa, we would end up with Schmidhuber's Model from 1990 (Schmidhuber, 1990) (**Figure 3**) for which Schmidhuber found evidence to develop self representation (Schmidhuber, 1991).

Interestingly, Damasio's descriptions follow a similar line in Damasio (1999). We depict a model implementing Damasio's core consciousness in **Figure 4G**. As Schmidhuber, Damasio requires a connection from the world model $\mathbf{w_t}$ to the body control system $\mathbf{h_t}$. However, in his view, consciousness does not emerge by itself. It is enforced by a "feeling" that is expressed as a training loss in terms of machine learning. As such, the Damasio model of core consciousness requires a loss that aims at the recovery of the image of the self in the world model. If this is implemented as a loss, we are able to express the desire to exist in the world. If implemented merely as trainable weights, we arrive at the theory of integrated information (Tononi et al., 2016) that creates consciousness as a maximally compressed representation of the world, the self, and their mutual interactions. Interestingly, these considerations also allow the integration of an attention mechanism (Vaswani et al., 2017) and other concepts of resolving context information used in machine learning. Realized in a biological learning framework, e.g., using neuromodulators like dopamine (Russek et al., 2017), the different notions of loss and trainable connections will disappear. Therefore, we hypothesize that from a meta-perspective, the models of Damasio (Damasio, 1999), Schmidhuber (Schmidhuber, 1990), Tononi (Tononi et al., 2016), Koch (Koch et al., 2016), and Dehaene (Dehaene et al., 1998, 2011, 2014; Dehaene and Naccache, 2001; Dehaene and Changeux, 2004; Sergent and Dehaene, 2004) may be basically regarded as different descriptions of the same fundamental principles.

Note that the models of consciousness that we have discussed so far are very basic. They do not take into account higher cognitive functions like language, different kinds of memory (procedural, episodic, semantic), nor any other complex multi-modal forms of processing, e.g., hierarchical action planning, induction, causal inference, or conclusion by analogy. Again, we follow Damasio at this point in **Figure 4H** in which all of these sophisticated processes are mapped into a single block "Memory / Extended Functions." Note, although we omit these extended functions, we are able to integrate them using trainable paths. As such, the model of core consciousness (**Figure 4G**) acts as a kind of"neural operating system" that is able to update and integrate also higher order cognitive functions according to the needs of the environment. We agree with Damasio that this core consciousness is shared by many species, i.e., probably all vertebrates, cephalopods, and perhaps even insects. By increasing the number of "extended functions," the degree of complexity and "integrated information" rises measurably, as also observed by Casarotto et al. (2016). In mammals, all higher order, i.e., extended, cognitive functions are located in the cerebral cortex. Hence, the growth of cortex size during mammal evolution, corresponds to an increasing number of extended functions.

This brings us back to the original heading of our section: There are clearly theories that enable modeling and implementation of consciousness in the machine. On the one hand, they are mechanistic to the extend that they can be implemented in programming languages and require similar inputs as humans would do. On the other hand, even the simple models in **Figure 4** are already arbitrarily complex, as every dashed path in the models could be realized by a deep neural network comprising many different layers. As such also training will be hard. Interestingly, the models follow a bottom-up strategy such that training and development can be performed in analogy to biological development and evolution. The models can be trained and grown to more complex tasks gradually.

6. DISCUSSION

Existence of consciousness in the machine is a hot topic of debate. Even with respect to the simple core consciousness, we observe opinions ranging from "generally impossible" (Carter et al., 2018) through "plausible" (Dehaene et al., 2017) to "has already been done" (Schmidhuber, 1991). Obviously, all of the suggested models cannot solve the qualia problem or the general problem on how to demonstrate whether a system is truly conscious. All of the emerging systems could merely be mimicking conscious behavior without being conscious at all (even **Figure 4A**). Yet as already discussed by Schmidhuber (1991), we would be able to measure correlates of self recognition similar to neural correlates of consciousness in humans (Koch et al., 2016) which could help to understand consciousness in human beings. However, as long as we have not solved how to provide proof of consciousness in human beings, we will also fail to do so in machines as the experience of consciousness is merely subjective.

Koch and Dehaene discussed the theories of global work space and integrated information as being opposed to each other (Carter et al., 2018). In the models found in **Figure 4**, we see that both concepts require a strong degree of interconnection. As such, we do not see why both concepts are fundamentally opposing. A global work space does not necessarily have to be encoded in decompressed state. Also, Maguire's view of integrated information (Maguire et al., 2016) is not necessarily impossible to implement mechanistically, as we are able to use concepts of deep learning to train highly integrated processing networks. In fact, as observed by neuroscience (Kriegeskorte and Douglas, 2018), both approaches might support each other yielding methods to construct and reproduce biological processes in a modular way. This allows the integration of representation (Gershman et al., 2015) and processing theories (Sun and Franklin, 2007) as long as they can be represented in terms of deep learning compatible operations (Maier et al., 2019b).

In all theories that we touched in this article, the notion of self is fundamental and the emergence of consciousness crucially requires embodiment. Feedback from internal body states is regarded to be the basis of emotions and feelings. Without emotions and feelings, the system cannot be trained and thus cannot adapt to new environments and changes of circumstances. Furthermore, certain additional cognitive functions are crucial to support, together with core consciousness, the emergence of extended consciousness as observed in humans and other non-human primates, as well as some other higher mammals, birds and cephalopods. These cognitive functions comprise, for instance attention, hierarchical action planning, procedural, episodic, and semantic memory.

In the machine learning inspired models, we assume that a disconnection between environment and self would cause a degradation of the system similar to the one that is observed in human beings in complete locked-in state (Kübler and Birbaumer, 2008) or in chronically curarized rats (Birbaumer, 2006). This homeostatsis, i.e., the regulation of body states aimed at maintaining conditions compatible with life, was also deemed important for the design of feeling machines by Man and Damasio (2019). Note that, a slightly different concept of homeostasis has been introduced by Tononi and Cirelli in the context of the sleep homeostasis hypothesis (Tononi and Cirelli, 2003). There, it is assumed that synaptic potentiation is tied to the homeostatic regulation of slow-wave activity.

Similar to the problems identified by Nagel, also the proposed mechanistic machine learning models will not be able to understand "what it is like" to be a bat. However, the notion of train-/learnable programs and connections or adapters might offer a solution to explore this in the future. Analogously, one cannot describe to somebody "what it is like" to play the piano or to snowboard on expert level unless one really acquires the ability. As such also the qualia problem persists in machine consciousness. However, we are able to investigate the actual configuration of the representation in the artificial neural net offering entirely new levels of insight.

In Damasio's theory, consciousness is effectively created by a training loss that causes the system to "want" to be conscious, i.e., "Cogito ergo sum" becomes "Sentio ergo sum." Comparison between trainable connections after (Schmidhuber, 1990), attention mechanisms (Vaswani et al., 2017), and this approach are within the reach of future machine learning models which will create new evidence for the discussion of integrated information and global work spaces. In fact, Schmidhuber has already taken up the work on combination of his early ideas with modern approaches from deep learning (Schmidhuber, 2015, 2018).

With models for extended consciousness, even the notion of the Homunculus (Kenny, 2016) can be represented by extension of the self with another self pointer. In contrast to common rejection of the Homunculus thought experiment, this recurrent approach can be trained using end-to-end systems comparable to AlphaGo (Silver et al., 2016).

Damasio also presents more interesting and important work that is mostly omitted in this article for brevity. In Damasio (1999), he also relates structural brain damage to functional loss of cognitive and conscious processing. Also the notion of emotion is crucial in a biological sense and is the driving effect of homeostasis. In Man and Damasio (2019), Damasio already pointed out that this concept will be fundamental for self-regulating robotic approaches.

With the ideas of cognitive computational neuroscience (Kriegeskorte and Douglas, 2018) and the approaches detailed above, we will design artificial systems that approach the mechanisms of biological systems in an iterative manner. With the iterations, the artificial systems will increase in complexity and similarity to the biological systems. With respect to artificial systems and machine learning, we are far away from the complexity of biological neural structures. Yet, we can adopt Damasio's strategy of identifying the presence of certain structures and links within these models. This is also the main contribution of our article. It allows us to link AI methods to Damasio's categories. In our analysis, we also observe that none of the machine learning models today could be mapped to the highest (and human-like) kind of extended consciousness.

However, even if we arrive at an artificial system that performs identical computations and reveals identical behavior as the biological system, we will not be able to deem this system as conscious beyond any doubts. The true challenge in being perceived as conscious will be the acceptance by human beings and society. Hence, requirements for conscious machines will comprise the similarity to biological conscious processes, the ability to convince human beings, and even the machine itself. As Alan Turing already proposed in his *imitation game* in 1950 to decide whether a machine is intelligent or even conscious (Turing, 1950), the ascription of such by other humans is a critical factor. For this purpose, Turing's Test has already been extended to also account for embodiment (French, 2012). However, such tests are only necessary, but not sufficient as Gary Marcus pointed out: Rather simple chat bot models are already able to beat Turing's Test in some occasions (Vardi, 2014).

Turing's test is able to test consciousness and intelligence only from an outside point of view. For the perception of consciousness, the internal point of view, however, is even more important. As Turing observed, the only means to quantify consciousness would be to directly compare experiences or memory contents with each other, or at least indirectly through serializations/projections. One way to serialize thoughts, experiences, and memories is language, and the comparison of such with real human serializations is part of the Turing-Test. However, since language provides only a coarse projection of memory and experience, also its analysis is necessarily coarse and may be feigned on purpose. A better path of creating a quantitative measure of consciousness would be to compare digital serializations of memory with respect to the ability to re-create the conscious experience with sufficient precision in the sense of the Nyquist-Shannon Sampling Theorem (Shannon, 1949). However, this path is still far away in the future as it would require storing and loading of digital memories using neural interfaces.

In this line of thought, we can now also relate to analysis of consciousness in biology: For example, there is the so called *mark and mirror test* (Bard et al., 2006) that shows at what

age self-representation appears in humans, or whether animals (corvids) seem to possess this ability. Here, the means of serialization of conscious experience is even more limited. Yet, we are able to understand certain basic feelings such as ones related to self-perception in the mirror by analysis of actions. So, one could interpret the mark and mirror test as a very weak version of the Turing test with respect to core consciousness. Another conjecture to assess the putative existence of artificial consciousness in candidate machine learning systems would be to apply the concept of the *perturbational complexity index* (Casali et al., 2013) which measures the degree of consciousness even in comatose or locked-in state patients.

Given the complexity and importance of the topic, we deem it necessary to look also at some ethical implications at this point.

6.1. Ethical Implications

Being able to create artificial systems that are indistinguishable from natural conscious beings and thus are also potentially conscious raises ethical concerns. First and foremost, in the transformation from core consciousness to extended consciousness, the systems gain the ability to link new program routines. As such systems followings such a line of implementation need to be handled with care and should be experimented on in a contained environment. With the right choice of embodiment in a virtual machine or in a robotic body, one should be able to solve such problems.

Of course there are also other ethical concerns, the more we approach human-like behavior. A first set of robotic laws has been introduced in Asimov's novels (Clarke, 1993). Even Asimov considered the rules problematic as can be seen from the plot twists in his novels. Aside this, being able to follow the robotic laws requires the robot to understand the concepts of "humans," "harm," and "self." Hence, such beings must be conscious. Therefore, tampering with their memories, emotions, and feelings is also problematic by itself. Being able to copy and reproduce the same body and mind does not lead to further simplification of the issue and implies the problem that we have to agree on ethics and standards of AI soon (Jobin et al., 2019).

7. CONCLUSION

In this article, we reviewed the state-of-the-art theories on consciousness in philosophy, neuroscience, AI, and machine learning. We find that the different disciplines need to interact to push research in this direction further. Interestingly, basic theories of consciousness can be implemented in computer programs. In particular, deep learning approaches are interesting as they offer the ability to train deep approximators that are not yet well-understood to construct mechanistic systems of complex neural and cognitive processes. We reviewed several machine learning architectures and related them to theories of strong reductionism and found that there are neural network architectures from which base consciousness could emerge. Yet, there is still a long way to form human-like extended consciousness.

DATA AVAILABILITY STATEMENT

The original contributions generated for the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation): grant KR5148/2-1 to PK – project number 436456810, the Emerging Talents Initiative (ETI) of the University Erlangen-Nuremberg (grant 2019/2-Phil-01 to PK). Furthermore, the research leading to these results has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (ERC grant no. 810316 to AM).

REFERENCES

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., et al. (2016). "Tensorflow: a system for large-scale machine learning," in *12th USENIX symposium on operating systems design and implementation (OSDI 16)*, (Savannah, GA), 265–283.

Baars, B. J. (1994). "A global workspace theory of conscious experience," in *Consciousness in Philosophy and Cognitive Neuroscience* (HOVE: Psychology Press), 149–171.

Baars, B. J. (2007). "The global workspace theory of consciousness," in *The Blackwell Companion to Consciousness*, eds S. Schneider and M. Velmans (Hoboken, NJ: Wiley-Blackwell), 236–246. doi: 10.1002/9780470751466.ch19

Baars, B. J., and Newman, J. (1994). "A neurobiological interpretation of global workspace theory," in *Consciousness in Philosophy and Cognitive Neuroscience* (Hove: Psychology Press), 211–226.

Banino, A., Barry, C., Uria, B., Blundell, C., Lillicrap, T., Mirowski, P., et al. (2018). Vector-based navigation using grid-like representations in artificial agents. *Nature* 557, 429–433. doi: 10.1038/s41586-018-0102-6

Bard, K. A., Todd, B. K., Bernier, C., Love, J., and Leavens, D. A. (2006). Self-awareness in human and chimpanzee infants: what is measured and what is meant by the mark and mirror test? *Infancy* 9, 191–219. doi: 10.1207/s15327078in0902_6

Barrett, D. G., Morcos, A. S., and Macke, J. H. (2019). Analyzing biological and artificial neural networks: challenges with opportunities for synergy? *Curr. Opin. Neurobiol.* 55, 55–64. doi: 10.1016/j.conb.2019.01.007

Benjamin, A. S., Qiu, C., Zhang, L. Q., Kording, K. P., and Stocker, A. A. (2019). "Shared visual illusions between humans and artificial neural networks," in *Proceedings of the Annual Conference of Cognitive Computational Neuroscience*. Available online at: https://ccneuro.org/2019/proceedings/0000585.pdf

Berger, H. (1934). Über das Elektrenkephalogramm des Menschen. *Deutsche Medizinische Wochenschrift* 60, 1947–1949. doi: 10.1055/s-0028-1130334

Birbaumer, N. (2006). Breaking the silence: brain-computer interfaces (BCI) for communication and motor control. *Psychophysiology* 43, 517–532. doi: 10.1111/j.1469-8986.2006.00456.x

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Heidelberg: Springer.

Boly, M., Massimini, M., Tsuchiya, N., Postle, B. R., Koch, C., and Tononi, G. (2017). Are the neural correlates of consciousness in the front or in the back of the cerebral cortex? Clinical and neuroimaging evidence. *J. Neurosci.* 37, 9603–9613. doi: 10.1523/JNEUROSCI.3218-16.2017

Boly, M., Seth, A. K., Wilke, M., Ingmundson, P., Baars, B., Laureys, S., et al. (2013). Consciousness in humans and non-human animals: recent advances and future directions. *Front. Psychol.* 4:625. doi: 10.3389/fpsyg.2013.00625

Brunette, E. S., Flemmer, R. C., and Flemmer, C. L. (2009). "A review of artificial intelligence," in *2009 4th International Conference on Autonomous Robots and Agents* (Wellington: IEEE), 385–392. doi: 10.1109/ICARA.2000.4804025

Campbell, K. (1992). *Body and Mind*. Notre Dame, IN: University of Notre Dame Press.

Carreira-Perpinan, M. A., and Hinton, G. E. (2005). "On contrastive divergence learning," in *Aistats* (Bridgetown: Citeseer), 33–40.

Carter, O., Hohwy, J., Van Boxtel, J., Lamme, V., Block, N., Koch, C., et al. (2018). Conscious machines: defining questions. *Science* 359, 400–400. doi: 10.1126/science.aar4163

Casali, A. G., Gosseries, O., Rosanova, M., Boly, M., Sarasso, S., Casali, K. R., et al. (2013). A theoretically based index of consciousness independent of sensory processing and behavior. *Sci. Transl. Med.* 5:198ra105. doi: 10.1126/scitranslmed.3006294

Casarotto, S., Comanducci, A., Rosanova, M., Sarasso, S., Fecchio, M., Napolitani, M., et al. (2016). Stratification of unresponsive patients by an independently validated index of brain complexity. *Ann. Neurol.* 80, 718–729. doi: 10.1002/ana.24779

Chalmers, D. J. (1995). Facing up to the problem of consciousness. *J. Conscious. Stud.* 2, 200–219.

Chauvin, Y., and Rumelhart, D. E. (1995). *Backpropagation: Theory, Architectures, and Applications*. Hove: Psychology Press.

Cho, K., Van Merriënboer, B., Bahdanau, D., and Bengio, Y. (2014). On the properties of neural machine translation: encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*. doi: 10.3115/v1/W14-4012

Chollet, F. (2015). *Keras*. Available online at: https://keras.io

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., and Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Sci. Rep.* 6:27755. doi: 10.1038/srep27755

Clarke, R. (1993). Asimov's laws of robotics: implications for information technology-Part I. *Computer* 26, 53–61. doi: 10.1109/2.247652

Cleeremans, A. (2005). Computational correlates of consciousness. *Prog. Brain Res.* 150, 81–98. doi: 10.1016/S0079-6123(05)50007-4

Crane, T. (2012). "The origins of qualia," in *History of the Mind-Body Problem*, eds T. Crane and S. Patterson (London: Routledge), 177–202. doi: 10.4324/9780203471029

Crick, F., and Koch, C. (1990). "Towards a neurobiological theory of consciousness," in *Seminars in the Neurosciences*, Vol. 2. (Amsterdam: ScienceDirect Elsevier), 263–275.

Crick, F., and Koch, C. (2003). A framework for consciousness. *Nat. Neurosci.* 6, 119–126. doi: 10.1038/nn0203-119

Cueva, C. J., and Wei, X.-X. (2018). Emergence of grid-like representations by training recurrent neural networks to perform spatial localization. *arXiv preprint arXiv:1803.07770*.

Damasio, A. (2001). Fundamental feelings. *Nature* 413, 781–781. doi: 10.1038/35101669

Damasio, A., and Meyer, K. (2009). "Consciousness: an overview of the phenomenon and of its possible neural basis," in *The Neurology of Consciousness: Cognitive Neuroscience and Neuropathology*, eds S. Laureys, and G. Tononi (Amsterdam: ScienceDirect Elsevier), 3–14. doi: 10.1016/B978-0-12-374168-4.00001-0

Damasio, A. R. (1999). *The Feeling of What Happens: Body and Emotion in the Making of Consciousness*. Boston, MA: Houghton Mifflin Harcourt.

Dehaene, S., and Changeux, J.-P. (2004). Neural mechanisms for access to consciousness. *Cogn. Neurosci.* 3, 1145–1158. Available online at: http://mathbrain.scicog.fr/actes_pdf/conf/dehaene_u1.pdf

Dehaene, S., Changeux, J.-P., and Naccache, L. (2011). "The global neuronal workspace model of conscious access: from neuronal architectures to clinical applications," in *Characterizing Consciousness: From Cognition to the Clinic?*, eds S. Dehaene and Y. Christen (Heidelberg: Springer), 55–84 doi: 10.1007/978-3-642-18015-6_4

Dehaene, S., Charles, L., King, J.-R., and Marti, S. (2014). Toward a computational theory of conscious processing. *Curr. Opin. Neurobiol.* 25, 76–84. doi: 10.1016/j.conb.2013.12.005

Dehaene, S., Kerszberg, M., and Changeux, J.-P. (1998). A neuronal model of a global workspace in effortful cognitive tasks. *Proc. Natl. Acad. Sci. U.S.A.* 95, 14529–14534. doi: 10.1073/pnas.95.24.14529

Dehaene, S., Lau, H., and Kouider, S. (2017). What is consciousness, and could machines have it? *Science* 358, 486–492. doi: 10.1126/science.aan8871

Dehaene, S., and Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. *Cognition* 79, 1–37. doi: 10.1016/S0010-0277(00)00123-2

Descartes, R. (1990). *Meditations on First Philosophy/Meditationes de Prima Philosophia: A Bilingual Edition*. Notre Dame, IN: University of Notre Dame Press. doi: 10.2307/j.ctvpj78hx

Engel, A. K., König, P., Kreiter, A. K., and Singer, W. (1991). Interhemispheric synchronization of oscillatory neuronal responses in cat visual cortex. *Science* 252, 1177–1179. doi: 10.1126/science.252.5009.1177

Franklin, S., and Graesser, A. (1999). A software agent model of consciousness. *Conscious. Cogn.* 8, 285–301. doi: 10.1006/ccog.1999.0391

French, R. M. (2012). Moving beyond the turing test. *Commun. ACM* 55, 74–77. doi: 10.1145/2380656.2380674

Gazzaniga, M. S. (2005). Forty-five years of split-brain research and still going strong. *Nat. Rev. Neurosci.* 6, 653–659. doi: 10.1038/nrn1723

Gazzaniga, M. S., Bogen, J. E., and Sperry, R. W. (1965). Observations on visual perception after disconnexion of the cerebral hemispheres in man. *Brain* 88, 221–236. doi: 10.1093/brain/88.2.221

Geerts, J. P., Stachenfeld, K. L., and Burgess, N. (2019). Probabilistic successor representations with kalman temporal differences. *arXiv preprint arXiv:1910.02532*. doi: 10.32470/CCN.2019.1323-0

Gershman, S. J. (2018). The successor representation: its computational logic and neural substrates. *J. Neurosci.* 38, 7193–7200. doi: 10.1523/JNEUROSCI.0151-18.2018

Gershman, S. J., Horvitz, E. J., and Tenenbaum, J. B. (2015). Computational rationality: a converging paradigm for intelligence in brains, minds, and machines. *Science* 349, 273–278. doi: 10.1126/science.aac6076

Gomez-Villa, A., Martín, A., Vazquez-Corral, J., and Bertalmío, M. (2018). Convolutional neural networks deceived by visual illusions. *arXiv [Preprint]* arXiv:1811.10565.

Gray, C. M., König, P., Engel, A. K., and Singer, W. (1989). Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties. *Nature* 338, 334–337. doi: 10.1038/338334a0

Grossberg, S. (2017). Towards solving the hard problem of consciousness: the varieties of brain resonances and the conscious experiences that they support. *Neural Netw.* 87, 38–95. doi: 10.1016/j.neunet.2016.11.003

Hafting, T., Fyhn, M., Molden, S., Moser, M.-B., and Moser, E. I. (2005). Microstructure of a spatial map in the entorhinal cortex. *Nature* 436, 801–806. doi: 10.1038/nature03721

Hagmann, P., Cammoun, L., Gigandet, X., Meuli, R., Honey, C. J., Wedeen, V. J., et al. (2008). Mapping the structural core of human cerebral cortex. *PLoS Biol.* 6:e159. doi: 10.1371/journal.pbio.0060159

Hameroff, S. (2001). "Biological feasibility of quantum approaches to consciousness," in *The Physical Nature of Consciousness* ed P. R. Van Loocke (Amsterdam: John Benjamins), 1–61. doi: 10.1075/aicr.29.02ham

Hameroff, S., and Penrose, R. (1996). Orchestrated reduction of quantum coherence in brain microtubules: a model for consciousness. *Math. Comput. Simul.* 40, 453–480. doi: 10.1016/0378-4754(96)80476-9

Hameroff, S., and Penrose, R. (2014). Consciousness in the universe: a review of the 'orch or' theory. *Phys. Life Rev.* 11, 39–78. doi: 10.1016/j.plrev.2013.08.002

Hassabis, D., Kumaran, D., Summerfield, C., and Botvinick, M. (2017). Neuroscience-inspired artificial intelligence. *Neuron* 95, 245–258. doi: 10.1016/j.neuron.2017.06.011

Hebb, D. O. (1949). The organization of behavior. na.

Herculano-Houzel, S. (2009). The human brain in numbers: a linearly scaled-up primate brain. *Front. Hum. Neurosci.* 3:31. doi: 10.3389/neuro.09.031.2009

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.* 9, 1735–1780. doi: 10.1162/neco.1997.9.8.1735

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Netw*. 4, 251–257. doi: 10.1016/0893-6080(91)90009-T

Hutter, M. (2004). *Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability*. Berlin: Springer Science & Business Media.

Jobin, A., Ienca, M., and Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nat. Mach. Intell.* 1, 389–399. doi: 10.1038/s42256-019-0088-2

Kaiser, L., Babaeizadeh, M., Milos, P., Osinski, B., Campbell, R. H., Czechowski, K., et al. (2019). Model-based reinforcement learning for atari. *arXiv preprint arXiv:1903.00374*.

Kenny, A. (2016). "The homunculus fallacy," in *Investigating Psychology*, ed J. Hyman (London: Routledge Taylor and Francis), 169–179.

Koch, C., and Hepp, K. (2006). Quantum mechanics in the brain. *Nature* 440, 611–611. doi: 10.1038/440611a

Koch, C., and Hepp, K. (2007). *The Relation Between Quantum Mechanics and Higher Brain Functions: Lessons from Quantum Computation and Neurobiology*. Citeseer. Available online at: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.652.712&rep=rep1&type=pdf

Koch, C., Massimini, M., Boly, M., and Tononi, G. (2016). Neural correlates of consciousness: progress and problems. *Nat. Rev. Neurosci*. 17:307. doi: 10.1038/nrn.2016.22

Kotov, A. A. (2017). A computational model of consciousness for artificial emotional agents. *Psychol. Russia State Art* 10, 57–73. doi: 10.11621/pir.2017.0304

Krauss, P., Metzner, C., Schilling, A., Tziridis, K., Traxdorf, M., Wollbrink, A., et al. (2018a). A statistical method for analyzing and comparing spatiotemporal cortical activation patterns. *Sci. Rep*. 8, 1–9. doi: 10.1038/s41598-018-23765-w

Krauss, P., Schilling, A., Bauer, J., Tziridis, K., Metzner, C., Schulze, H., et al. (2018b). Analysis of multichannel EEG patterns during human sleep: a novel approach. *Front. Hum. Neurosci*. 12:121. doi: 10.3389/fnhum.2018.00121

Kriegeskorte, N., and Douglas, P. K. (2018). Cognitive computational neuroscience. *Nat. Neurosci*. 21, 1148–1160. doi: 10.1038/s41593-018-0210-5

Kübler, A., and Birbaumer, N. (2008). Brain-computer interfaces and communication in paralysis: extinction of goal directed thinking in completely paralysed patients? *Clin. Neurophysiol*. 119, 2658–2666. doi: 10.1016/j.clinph.2008.06.019

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. (2017). Building machines that learn and think like people. *Behav. Brain Sci*. 40. doi: 10.1017/S0140525X16001837

Lange, S., and Riedmiller, M. (2010). "Deep auto-encoder neural networks in reinforcement learning," in *The 2010 International Joint Conference on Neural Networks (IJCNN)* (Barcelona: IEEE), 1–8. doi: 10.1109/IJCNN.2010.5596468

Laplane, L., Mantovani, P., Adolphs, R., Chang, H., Mantovani, A., McFall-Ngai, M., et al. (2019). Opinion: why science needs philosophy. *Proc. Natl. Acad. Sci. U.S.A*. 116, 3948–3952. doi: 10.1073/pnas.1900357116

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

Levine, J. (2001). *Purple Haze: The Puzzle of Consciousness*. Oxford: Oxford University Press. doi: 10.1093/0195132351.001.0001

Libet, B. (1985). Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behav. Brain Sci*. 8, 529–539. doi: 10.1017/S0140525X00044903

Libet, B., Wright, E. W. Jr., and Gleason, C. A. (1983). Preparation- or intention-to-act, in relation to pre-event potentials recorded at the vertex. *Electroencephalogr. Clin. Neurophysiol*. 56, 367–372. doi: 10.1016/0013-4694(83)90262-6

Maguire, P., Moser, P., and Maguire, R. (2016). Understanding consciousness as data compression. *J. Cogn. Sci*. 17, 63–94. doi: 10.17791/jcs.2016.17.1.63

Maier, A., Steidl, S., Christlein, V., and Hornegger, J. (2018). *Medical Imaging Systems: An Introductory Guide*. Heidelberg: Springer. doi: 10.1007/978-3-319-96520-8

Maier, A., Syben, C., Lasser, T., and Riess, C. (2019a). A gentle introduction to deep learning in medical image processing. *Zeitschrift für Medizinische Physik* 29, 86–101. doi: 10.1016/j.zemedi.2018.12.003

Maier, A. K., Syben, C., Stimpel, B., Würfl, T., Hoffmann, M., Schebesch, F., et al. (2019b). Learning with known operators reduces maximum error bounds. *Nat. Mach. Intell*. 1, 373–380. doi: 10.1038/s42256-019-0077-5

Man, K., and Damasio, A. (2019). Homeostasis and soft robotics in the design of feeling machines. *Nat. Mach. Intell*. 1, 446–452. doi: 10.1038/s42256-019-0103-7

Mao, J., Gan, C., Kohli, P., Tenenbaum, J. B., and Wu, J. (2019). The neuro-symbolic concept learner: interpreting scenes, words, and sentences from natural supervision.

Marblestone, A. H., Wayne, G., and Kording, K. P. (2016). Toward an integration of deep learning and neuroscience. *Front. Comput. Neurosci*. 10:94. doi: 10.3389/fncom.2016.00094

Minsky, M., and Papert, S. A. (2017). *Perceptrons: An Introduction to Computational Geometry*. Cambridge, MA: MIT Press. doi: 10.7551/mitpress/11301.001.0001

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., et al. (2015). Human-level control through deep reinforcement learning. *Nature* 518, 529–533. doi: 10.1038/nature14236

Mohsenzadeh, Y., Mullin, C., Lahner, B., and Oliva, A. (2020). Emergence of visual center-periphery spatial organization in deep convolutional neural networks. *Sci. Rep*. 10, 1–8. doi: 10.1038/s41598-020-61409-0

Nagel, T. (1974). What is it like to be a bat? *Philos. Rev*. 83, 435–450.

Nasr, K., Viswanathan, P., and Nieder, A. (2019). Number detectors spontaneously emerge in a deep neural network designed for visual object recognition. *Sci. Adv*. 5:eaav7903. doi: 10.1126/sciadv.aav7903

Newman, J., and Baars, B. J. (1993). A neural attentional model for access to consciousness: a global workspace perspective. *Concepts Neurosci*. 4, 255–290.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). "Pytorch: an imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Vancouver, BC), 8026–8037.

Penrose, R. (1994). Mechanisms, microtubules and the mind. *J. Conscious. Stud*. 1, 241–249.

Penrose, R. (2001). Consciousness, the brain, and spacetime geometry: an addendum: some new developments on the orch or model for consciousness. *Ann. N. Y. Acad. Sci*. 929, 105–110. doi: 10.1111/j.1749-6632.2001.tb05710.x

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., et al. (2019). Machine behaviour. *Nature* 568, 477–486. doi: 10.1038/s41586-019-1138-y

Rees, G., Kreiman, G., and Koch, C. (2002). Neural correlates of consciousness in humans. *Nat. Rev. Neurosci*. 3, 261–270. doi: 10.1038/nrn783

Reggia, J. A., Katz, G., and Huang, D.-W. (2016). What are the computational correlates of consciousness? *Biol. Inspired Cogn. Arch*. 17, 101–113. doi: 10.1016/j.bica.2016.07.009

Rey, G. (1988). "A question about consciousness," in *Perspectives on Mind*, eds H.R. Otto, J. Tuedio (Heidelberg: Springer), 5–24 doi: 10.1007/978-94-009-4033-8_2

Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., et al. (2019). A deep learning framework for neuroscience. *Nat. Neurosci*. 22, 1761–1770. doi: 10.1038/s41593-019-0520-2

Rosen, S. M. (2015). Why natural science needs phenomenological philosophy. *Prog. Biophys. Mol. Biol*. 119, 257–269. doi: 10.1016/j.pbiomolbio.2015.06.008

Russek, E. M., Momennejad, I., Botvinick, M. M., Gershman, S. J., and Daw, N. D. (2017). Predictive representations can link model-based reinforcement learning to model-free mechanisms. *PLoS Comput. Biol*. 13:e1005768. doi: 10.1371/journal.pcbi.1005768

Savage, N. (2019). Marriage of mind and machine. *Nature* 571, 15–27. doi: 10.4324/9780429281662-2

Schmidhuber, J. (1990). "An on-line algorithm for dynamic reinforcement learning and planning in reactive environments," in *1990 IJCNN International Joint Conference on Neural Networks* (San Diego, CA: IEEE), 253–258. doi: 10.1109/IJCNN.1990.137723

Schmidhuber, J. (1991). "A possibility for implementing curiosity and boredom in model-building neural controllers," in *Proc. of the International Conference on Simulation of Adaptive Behavior: From Animals to Animats* (Paris), 222–227.

Schmidhuber, J. (2015). On learning to think: algorithmic information theory for novel combinations of reinforcement learning controllers and recurrent neural world models. *arXiv preprint arXiv:1511.09249*.

Schmidhuber, J. (2018). One big net for everything. *arXiv preprint arXiv:1802.08864*.

Searle, J. R. (1980). Minds, brains, and programs. *Behav. Brain Sci.* 3, 417–424. doi: 10.1017/S0140525X00005756

Sergent, C., and Dehaene, S. (2004). Neural processes underlying conscious perception: experimental findings and a global neuronal workspace framework. *J. Physiol.* 98, 374–384. doi: 10.1016/j.jphysparis.2005.09.006

Seth, A. (2009). Explanatory correlates of consciousness: theoretical and computational challenges. *Cogn. Comput.* 1, 50–63. doi: 10.1007/s12559-009-9007-x

Seth, A. K., Dienes, Z., Cleeremans, A., Overgaard, M., and Pessoa, L. (2008). Measuring consciousness: relating behavioural and neurophysiological approaches. *Trends Cogn. Sci.* 12, 314–321. doi: 10.1016/j.tics.2008.04.008

Shannon, C. E. (1949). Communication in the presence of noise. *Proc. IRE* 37, 10–21. doi: 10.1109/JRPROC.1949.232969

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., et al. (2016). Mastering the game of go with deep neural networks and tree search. *Nature* 529:484. doi: 10.1038/nature16961

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., et al. (2017). Mastering the game of go without human knowledge. *Nature* 550, 354–359. doi: 10.1038/nature24270

Singer, W. (1993). Synchronization of cortical activity and its putative role in information processing and learning. *Annu. Rev. Physiol.* 55, 349–374. doi: 10.1146/annurev.ph.55.030193.002025

Singer, W., and Gray, C. M. (1995). Visual feature integration and the temporal correlation hypothesis. *Annu. Rev. Neurosci.* 18, 555–586. doi: 10.1146/annurev.ne.18.030195.003011

Solomonoff, R. J. (1964). A formal theory of inductive inference. Part I. *Inform. Control* 7, 1–22. doi: 10.1016/S0019-9958(64)90223-2

Sporns, O., Tononi, G., and Kötter, R. (2005). The human connectome: a structural description of the human brain. *PLoS Comput. Biol.* 1:e42. doi: 10.1371/journal.pcbi.0010042

Srinath, R., Emonds, A., Wang, Q., Lempel, A. A., Dunn-Weiss, E., Connor, C. E., et al. (2020). Early emergence of solid shape coding in natural and deep network vision. *Curr. Biol.* 31, 1–15. doi: 10.1016/j.cub.2020.09.076

Stachenfeld, K. L., Botvinick, M. M., and Gershman, S. J. (2017). The hippocampus as a predictive map. *Nat. Neurosci.* 20:1643. doi: 10.1038/nn.4650

Starzyk, J. A., and Prasad, D. K. (2010). "Machine consciousness: a computational model," in *Brain-inspired Cognitive Systems (BICS 2010)*, eds C. Hernández, R. Sanz, J. Gómez Ramirez, L. S. Smith, A. Hussain, A. Chella, I. Aleksander (Heidelberg: Springer).

Sun, R., and Franklin, S. (2007). "'Computational models of consciousness: a taxonomy and some examples," in *The Cambridge Handbook of Consciousness (Cambridge Handbooks in Psychology)*, eds E. Thompson, M. Moscovitch, and P. D. Zelazo (Cambridge: Cambridge University Press).

Sutton, R. S., and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.

Tegmark, M. (2000). Importance of quantum decoherence in brain processes. *Phys. Rev. E* 61:4194. doi: 10.1103/PhysRevE.61.4194

Tegmark, M. (2014). Consciousness is a state of matter, like a solid or gas. *N. Sci.* 222, 28–31. doi: 10.1016/S0262-4079(14)60731-4

Tegmark, M. (2015). Consciousness as a state of matter. *Chaos Solitons Fractals* 76, 238–270. doi: 10.1016/j.chaos.2015.03.014

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., and Goodman, N. D. (2011). How to grow a mind: statistics, structure, and abstraction. *Science* 331, 1279–1285. doi: 10.1126/science.1192788

Thagard, P. (2009). Why cognitive science needs philosophy and vice versa. *Topics Cogn. Sci.* 1, 237–254. doi: 10.1111/j.1756-8765.2009.01016.x

Tononi, G. (2004). An information integration theory of consciousness. *BMC Neurosci.* 5:42. doi: 10.1186/1471-2202-5-42

Tononi, G. (2008). Consciousness as integrated information: a provisional manifesto. *Biol. Bull.* 215, 216–242. doi: 10.2307/25470707

Tononi, G., Boly, M., Massimini, M., and Koch, C. (2016). Integrated information theory: from consciousness to its physical substrate. *Nat. Rev. Neurosci.* 17, 450–461. doi: 10.1038/nrn.2016.44

Tononi, G., and Cirelli, C. (2003). Sleep and synaptic homeostasis: a hypothesis. *Brain Research Bull.* 62, 143–150. doi: 10.1016/j.brainresbull.2003.09.004

Tononi, G., and Koch, C. (2008). The neural correlates of consciousness-an update. *Ann. N. Y. Acad. Sci.* 1124, 239–261. doi: 10.1196/annals.1440.004

Tsoi, A. C., and Back, A. (1997). Discrete time recurrent neural network architectures: a unifying review. *Neurocomputing* 15, 183–223. doi: 10.1016/S0925-2312(97)00161-6

Turing, A. (1950). Computing machinery and intelligence. *Mind* 59:433. doi: 10.1093/mind/LIX.236.433

Van Gerven, M. (2017). Computational foundations of natural intelligence. *Front. Comput. Neurosci.* 11:112. doi: 10.3389/fncom.2017.00112

van Hateren, J. (2019). A theory of consciousness: computation, algorithm, and neurobiological realization. *Biol. Cybernet.*113, 357–372. doi: 10.1007/s00422-019-00803-y

Vardi, M. Y. (2014). Would turing have passed the turing test? *Commun. ACM* 57, 5–5. doi: 10.1145/2643596

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention is all you need," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Vancouver, BC), 5998–6008.

Wang, Y., Yao, H., and Zhao, S. (2016). Auto-encoder based dimensionality reduction. *Neurocomputing* 184, 232–242. doi: 10.1016/j.neucom.2015.08.104

Watanabe, E., Kitaoka, A., Sakamoto, K., Yasugi, M., and Tanaka, K. (2018). Illusory motion reproduced by deep neural networks trained for prediction. *Front. Psychol.* 9, 345. doi: 10.3389/fpsyg.2018.00345

Weisberg, J. (2020). *The Hard Problem of Consciousness*. Available online at: https://www.iep.utm.edu/hard-con/ (accessed February 28, 2020).

Weiskrantz, L., and Warrington, E. (1975). Blindsight-residual vision following occipital lesions in man and monkey. *Brain Res.* 85, 184–185. doi: 10.1016/0006-8993(75)91036-7

Weiskrantz, L., Warrington, E. K., Sanders, M., and Marshall, J. (1974). Visual capacity in the hemianopic field following a restricted occipital ablation. *Brain* 97, 709–728. doi: 10.1093/brain/97.1.709

Young, T., Hazarika, D., Poria, S., and Cambria, E. (2018). Recent trends in deep learning based natural language processing. *IEEE Comput. Intell. Mag.* 13, 55–75. doi: 10.1109/MCI.2018.2840738

Zador, A. M. (2019). A critique of pure learning and what artificial neural networks can learn from animal brains. *Nat. Commun.* 10, 1–7. doi: 10.1038/s41467-019-11786-6

Zeman, A. A., Ritchie, J. B., Bracci, S., and de Beeck, H. O. (2020). Orthogonal representations of object shape and category in deep convolutional neural networks and human visual cortex. *Sci. Rep.* 10, 1–12. doi: 10.1038/s41598-020-59175-0

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.