



ELSEVIER

Contents lists available at ScienceDirect

New Techno-Humanities

journal homepage: www.elsevier.com/locate/techum

Re-creating the world - On necessary features for the creation of AGI

Oliver Li¹

Researcher in the WASP-HS projects 'Artificial Intelligence, Democracy and Human Dignity' and 'The Artificial Public Servant', Sweden

ARTICLE INFO

Keywords:

AGI
Embodiment
Self-reflexivity
Creation of AGI
Mind

ABSTRACT

In this paper I identify and discuss a number of features that I argue are necessary for the realization of AGI. As a preliminary step, common definitions of AGI are presented in respect to their understanding of mind, intelligence, and consciousness. I show that, despite the amazing performance of artificial systems, at present they are still far from exhibiting AGI, and I identify some of their central short-comings. Secondly, inspired by research within the philosophy of mind, embodiment and situatedness, I suggest a number of features that I deem necessary for a mind. I then investigate the possible objection against the relevance of these features namely that they are overly anthropocentric or biocentric. I further discuss aspects of these features in relation to their transfer to artificial systems with the goal of creating an artificial mind. I finally conclude that self-reflexivity and the re-creation of the world as an inner world should be strongly focused upon if one wishes to create an artificial mind or artificial consciousness. However, I also issue a warning about some well-known risks when creating AGI.

1. Introduction

In the debate surrounding the development of artificial systems, terms including strong-AI, artificial general intelligence (AGI), and artificial super-intelligence (ASI) are frequently used to describe an artificial form of intelligence which at least matches and even surpasses the abilities associated with intelligence in humans (see, for example, Coeckelbergh 2020, 15; Mitchell 2019, 46, Kaplan and Haenlein 2019; Bostrom 2014). As I shall argue, cognitive abilities like having consciousness or, in other words, "having a mind" should presumably be included. An example of recent developments, which may give the *impression* that the achievement of such general artificial intelligence is near at hand, can be found in the OpenAI program and its advances in the program GPT (Generative Pretrained Transformer) (see, for example, Brown et al. 2020). OpenAI also actively and openly supports and strives for the development of AGI (Altman 2023) Although the achievements of these systems appear impressive, there are clear reasons to believe the contrary to be true (Floridi and Chiriatti 2020).

In this paper, I identify and discuss a number of features which I argue are necessary for the realization of AGI. First, I present common definitions of AGI with attention to their understanding of mind, intelligence, and consciousness. I show that, despite the amazing performance of artificial systems, they do not exhibit AGI at present, and I identify some of their central short-comings. Secondly, inspired by research within the philosophy of mind, the philosophy of AI related to embodiment and situatedness (for example Müller 2007; Lyre 2020;

Crosby 2020), I suggest a number of features that I deem necessary for a mind. I then investigate the possible objection against the relevance of these features: that they are overly anthropocentric or biocentric. Given the result that these features are *not* merely of relevance for humans and an understanding of the human mind, but that they are in some sense universal, I further discuss aspects of these features related to their transfer to artificial systems with the goal of creating an artificial mind. I finally conclude that self-reflexivity and the re-creation of the world as an inner world should be strongly focused upon if one wishes to create an artificial mind or artificial consciousness. However, I also issue a warning about some well-known risks of creating AGI.

1.1. Some preliminaries about the term AGI - artificial general intelligence

There are a variety of ways to define or describe what is meant by both Artificial Intelligence (AI) in general and Artificial General Intelligence (AGI) in particular. I have already introduced one possible very loose definition of AGI: namely an artificial form of intelligence which at least matches and even surpasses the abilities associated with intelligence in humans. However, what would such abilities be?

A possible definition of AGI, here denoted as strong-AI in contrast to weak-AI, the cognitive abilities would presumably include "having a mind" or, as I would interpret it, being conscious. "On the Strong AI view, the appropriately programmed digital computer does not just simulate having a mind, it literally has a mind." (Searle 2004, 46) Apparently, Searle thinks that having a mind would be the crucial feature

E-mail addresses: oliver.li@teol.uu.se, oliver.li@crs.uu.com

¹ Address: Department of Theology, Center for Multidisciplinary Research on Religion and Society (CRS, Uppsala), Uppsala University, Box 511 SE-75120 Uppsala, Sweden.

<https://doi.org/10.1016/j.techum.2023.05.004>

Received 10 October 2021; Received in revised form 24 May 2023; Accepted 26 May 2023

Available online xxx

2664-3294/© 2023 The Author(s). Published by Elsevier Ltd on behalf of Shanghai Jiao Tong University. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>)

for an AGI. I assume that having a mind, at least at a higher level, would include some form of subjectivity. Similarly, David J. Chalmers focuses on the machine's having mentality: "the field of *artificial intelligence* (or AI) is devoted in large part to the goal of reproducing mentality in computational machines." (Chalmers 1996, 313)

In her study of AI, Manuela Lenzen focuses on different approaches in the field of AI. Firstly, some researchers wish to develop systems that perform specific tasks. This is presumably the most widespread type of research in the field of AI and corresponds to weak-AI or AI applications in a great variety of areas in everyday life, such as speech- and face recognition or self-driving cars. Another area identified by Lenzen would be attempts to understand human cognition by understanding artificial systems. Finally, there are researchers who wish to develop AGI, in the sense that the machine has at least the flexibility and universality of the human mind (Lenzen 2018, 31–34).

Apparently, all of the above attempts to specify the term AGI are vague in some sense but nevertheless implicitly seem to include the following: consciousness, and universal cognitive abilities which at least are on the same level as in humans. If consciousness is on a human level this would presumably even include self-consciousness.

Also, in all of these forms of artificial intelligence, the term AI implies that some form of "intelligence" is involved. Here, I believe, it is essential to observe that intelligence *cannot be equated* with having consciousness or having a mind. It is easy to imagine, for example, humans who are fully conscious and self-conscious but are not intelligent in certain areas, as for example performing advanced mathematical operations. Likewise, it is easy to imagine machines that are highly intelligent in terms of performing certain tasks but are definitely *not* conscious.

Furthermore, the focus of the above tentative definitions importantly focus on the *cognitive* or *mental* abilities of the system in question and do not directly involve the body. Questions concerning the body in the discourse of artificial system are usually discussed in the context of robotics. However, it should be mentioned that although mental and cognitive abilities can be applied to abstract problems, these abilities—inherent in all biological systems—are trained and developed in interaction with the surroundings of the system. Such interaction is accomplished by some form of *body* in its broadest sense. Thus, since consciousness and cognitive or mental abilities in biological systems can be regarded as developing bottom-up from simple, rudimentary forms to complex forms, a similar view seems to be present in the assessment of the possibility of AGI in non-biological systems. Also, simply demarcating consciousness by appeal to *human*-consciousness might lead to the problems which follow speciesism, as Peter Singer has elaborated in relation to animal rights (Singer 2009). It is also far from clear *how* consciousness in a machine might manifest itself; that is, what it would be like to be a conscious machine (Seth 2021, 263). Thus, as shall become clear further down, not fully considering the role of the bodily or developmental aspects may turn out to be a fatal mistake. To be sure, body or bodily aspects may have to be understood in a very broad sense, not necessarily in terms of having the equivalency of arms, legs, ears, eyes, or the like.

In summary, the following tentative understanding of AGI will serve as a starting-point in the subsequent discussion: AGI is a system which has consciousness and universal cognitive abilities at least on the same level as humans.

1.2. GPT - there is nobody at home yet!

In terms of simple performance of specific tasks, contemporary advanced AI-systems, like the GPT, surely are impressive, but are they even close to becoming an AGI? Do they incorporate consciousness and universal cognitive abilities on the same level as humans? The following excerpt of a text is entirely generated by an implementation of this

program. The program responds to the question "What is love?" and produces a seemingly adequate answer¹:

Love is a feeling that people experience, which transcends the ability to reason. This kind of love can only be felt by humans, or rather, homo sapiens (and perhaps other beings) since we have the ability to feel emotions. Love is more than just an emotion; it must also involve rational thought in order for one to understand and perceive love fully. The brain has evolved over many generations through natural selection so as to give us the faculties required for us to feel and think about such things (Ayfer 2020).

In a second run the same program begins its answer with:

Love, from the human perspective, is a phenomenon that has evolved in humans over thousands of years. It is distinctly different from lust or desire but closely related to it. While there are many forms of love (romantic, platonic and familial), they all share common features (Ayfer 2020).

The authors of the web-page add an important comment and disclaimer and point out that GPT "lacks ... opinions or knowledge - it merely mimics opinions ..." (Ayfer 2020).²

Both texts about love could very well have been written by a human. Merely reading them does not yield hints that they are produced by an AI. Still, the disclaimer of the authors of the web-page sharing the answers already emphasizes that the AI—the GPT—*lacks knowledge* and "merely mimics" having knowledge. There is nobody at home. Also, the fact that both texts are significantly different, hints that they are not produced by a human. If I, for example, were to write something on a topic within an interval of a few minutes, I would presumably produce two texts which are significantly similar to each other. To be sure, consistency between consecutive texts could be achieved even without understanding. Nevertheless, the ability of GPT to produce coherent, consistent, and informative texts is, to say the least, *impressive*. If this ability could, and possibly were, incorporated in a system or developed into a system which *has* consciousness and self-consciousness, then surely GPT within that system would be a *very* powerful tool.

1.3. Shortcomings of AI-systems based on machine learning

But what does the GPT-3 lack and what would be needed to develop it into a system which might be conscious? In her enlightening and informative book *Artificial Intelligence*, computer scientist Melanie Mitchell highlights a number of shortcomings of present-day artificial systems, and important differences between the abilities of AI systems and corresponding human abilities (Mitchell 2019). These short-comings and differences will form one of the starting points for my suggestions of features that I deem necessary for a mind.

Many of the present AI systems are based on deep learning algorithms and variations of machine learning. These variations of machine learning can be regarded as the driving force in the current developments in artificial intelligence (Alpaydin 2016, xiii; Mitchell 2019,

¹ In mid-2020 the webpage Philosopher AI was available free of charge. As of July 2021 one has to *purchase* 10 queries at a time.

² The complete disclaimer of the authors of the web-page is as follows: "This is an experiment in what one might call "prompt engineering," which is a way to utilize GPT-3, a neural network trained and hosted by OpenAI. GPT-3 is a language model. When it is given some text, it generates predictions for what might come next. It is remarkably good at adapting to different contexts, as defined by a prompt (in this case, hidden), which sets the scene for what type of text will be generated. Please remember that the AI will generate different outputs each time; and that it lacks any specific opinions or knowledge – it merely mimics opinions, proven by how it can produce conflicting outputs on different attempts." (Ayfer 2020) As of 2023, there are upgraded version like GPT-3.5 and GPT-4. There are also similar AI-based chatbots developed by other companies than OpenAI, like Google Bard based on LaMDA. Presumably, these version will produce 'better' results. This will, however, not affect the overall reasoning in this article.

35–43). Machine learning is the more general term, and deep learning can be regarded as a subset of machine learning (Alpaydin 2016). Commonly, machine learning is based on artificial neural networks. Such neural networks consist of an input layer, an output layer, and one or more ‘hidden layers.’ In the case of deep learning or deep neural networks, several hidden layers are involved in the neural network. The ‘neurons’ or nodes in these networks are interconnected, and the interconnections are weighted such that a given input will produce the expected output (see, for example, Alyapadin 2016, chap 4).³ The weights of the interconnections are adjusted by a process called backpropagation, which calculates the error of the output and adjusts the weights of the interconnections accordingly (Alpaydin 2016, Mitchell 2019, 79–80). Furthermore, it is common to distinguish at least between supervised, unsupervised, and reinforcement learning. The first uses given categories both in the input and output datasets. Examples and non-examples for the categories in question are provided to the system. In the case of unsupervised learning, the algorithm itself establishes the categories. In reinforcement learning, the output is assessed by the programmer, who provides feedback on whether the output is ‘good’ or ‘bad’ (Coeckelbergh 2020, 83–87).

The process of deep learning can be regarded as a statistical analysis of the ‘learning data’ (Pasquenelli 2019, 10,17). By construction, this analysis is based on an input-output model. Some data is fed into the system, for example, a picture, and the system produces an output, for example, the category to which the picture belongs (Pasquenelli 2019, 7–8). This observation will be of relevance in the suggestions for the necessary conditions of an AGI.

Another interesting class of machine learning frameworks has been discussed by Dmytro Mykhailov and Nicola Liberati. They point out that certain classes of machine learning (Generative Adversarial Networks) do not simply ‘memorize’ data, but utilize two ‘competing’ networks, the generator and the discriminator and reason on the basis of the postphenomenological concept of ‘technological intentionality’ that “...computers are always more than what the programmer, the designer, and the user merely perceive.” (Mykhailov and Liberati 2022, 10–11).

It should be noted already at this stage that unsupervised learning presumably is closest to the kind of learning required for an AGI if it is to match universal cognitive abilities on the same level as humans. However, one should observe that ‘unsupervised’ does not mean ‘no involvement’ of humans but rather that the categories produced in the learning process are not fixed beforehand.

Mitchell’s analysis and description of artificial systems based on deep learning uncovers some serious issues, which have not yet been solved by computer scientists. A first observation is that, although the internal structure of deep learning algorithms is modeled to reflect the structure of a biological neuronal network, there are significant differences, in particular, regarding their performance and how they are trained. The most apparent difference in how deep learning AI-systems are trained compared to humans is reflected in the amount of training data. Whereas a child learns to recognize, for example, a cow by looking at a cow and being told that the cow is a cow *once*, a deep learning program requires big data (Mitchell 2019, 98; Watson 2019, 423). Thus deep learning systems are *inefficient*. They achieve their results by involving big data, and high computational power. It has also been shown that modeling the *full* functionality of one single biological neuron requires between five and eight layers of deep neural network with a depth of between 128–256 artificial neurons in each layer (Beniaguev, Segev, and London 2021).

Another problem and difference between human learning and machine learning is reflected in the possibility of “adversarial attacks”: a picture can be modified such that it in principle looks the same

to a human but the algorithm for recognizing or captioning the picture is tricked and produces a totally inadequate output. An AI-system could, for example, mistake a school-bus for an ostrich if the picture has been modified to fool the system. (Mitchell 2019, 110–15, 228–31; Watson 2019, 422–23, Pasquinelli 2019, 17)). This phenomenon is also known as brittleness; the system, although it may classify pictures with high accuracy, is easily fooled; it is brittle. Apparently this problem reflects the fact that AI-system do not *understand* what they see.

Often the problem of bias and the fact that AI systems work with approximations based on statistical models are mentioned as shortcomings of AI systems based on deep learning (Mitchell 2019, 106–108; Pasquenelli 2019, 10,13). Indeed, these problems can be seen as a reflection of the lack of semantic understanding in AI systems. However, one should be aware that also humans are subject to bias and constantly base their perceptions – and even decisions – on approximations.

Mitchell points out that it is not easy to defend the AI-system against adversarial attacks and that the ultimate issue behind this flaw - and even other problems like the above mentioned bias - is the question of *understanding* (Mitchell 2019, 114). She subsequently discusses what she denotes as the “barrier of meaning”, and describes efforts to solve the problem of creating AI-systems that have understanding, can make abstractions, have knowledge, have analogy-making abilities and so on (Mitchell 2019, 247–65), that is, that a system has universal cognitive abilities.

The problem of understanding, or the barrier of meaning, has also recently been addressed in attempts to incorporate common-sense knowledge in deep-learning (Bosselut et al. 2020; Sap et al. 2019). Such attempts are nevertheless based on big data, and pre-training of the system based on such data. Presumably, similar problems to the above mentioned bias would also occur in these systems. Furthermore, although the systems can produce knowledge that humans deem to be common-sense knowledge, the system is *not* embedded in a context in which the knowledge produced actually relates to some experienced reality, it uses *propositions* of what can be regarded as common-sense. At best, I believe, the data used in training allows for indirect grounding of the knowledge in question as Lyre has argued in relation to, for example, Google Translate (Lyre 2020, 337).

However, does Mitchell hint other possible paths for developing AGI in her analysis of the functionality and the deficiencies of AI-systems? In at least two points, she suggests in which direction researchers could think. Firstly, in her comparison of object recognition in the brain and convoluted neural networks, she observes that the information flow in the visual cortex occurs in *both* directions: from higher layers in the cortex to lower and vice versa. She points out that the backwards connections are not well understood by neuroscientists but that “[...] it is well established that our prior knowledge and expectations, presumably stored in higher brain layers, strongly influence what we perceive” (Mitchell 2019, 73). It seems that the visual cortex, in contrast to many AI-systems, is *not* merely a feed-forward or input-output system, as observed above, but that feedback from higher layers possibly plays an important role in how the visual cortex works in “recognizing” patterns. I deliberately put recognizing in quotations marks since the process of recognizing, for example, visual patterns, is *not* an isolated process restricted to the visual cortex but is integrated in the global activity of the individual’s brain. Her observation seems to suggest that feedback processes might play an important role in the development of AGI-systems.

Another possible path for the development of AGI is hinted at by Mitchell, who herself has made significant contributions to implementing analogy-making in computers (Mitchell 1993), when she writes: “But after grappling with AI for many years, I am finding the embodiment argument increasingly compelling” (Mitchell 2019, 265). According to her, the embodiment hypothesis is the following: “[A] machine cannot attain human-level intelligence without having some kind of body that interacts with world” (Mitchell 2019, 265). Mitchell introduces this hypothesis in her discussion of the “barrier of meaning”, and how un-

³ For a more technical and mathematical accurate description, see textbooks on deep learning, for example, Charniak, Eugen *Introduction to Deep Learning* (2018).

derstanding, abstraction, analogy-making, and other related cognitive abilities could be implemented in AI-systems.

The premise that embodiment, often linked to the closely-related premise of situatedness, (that is, at least in some sense, it is *necessary* for any system, be it biological or based on silicon, to have or host a mind), has been explored by other researchers in the closely-related fields of robotics, AI philosophy, and, unsurprisingly, in the philosophy of mind. (For discussions involving embodiment or situatedness, see: Crosby 2020; Lyre 2020; Baldassarre et al. 2017; Abramson 2011; Wilson and Foglia 2017; MacLennan 2017; Müller 2007). It can, in fact, be argued that embodiment entails situatedness and vice versa (Müller 2007, 105). One of the first attempts to implement ideas based on embodiment was made by Rodney A. Brooks who, already in 1991, successfully built robots able to operate without supervision (Brooks 1991). It has also been argued, on the grounds of postphenomenology, that the technological artifact can be regarded as an extension of the human body, thus an AI-system would be part of the human's body and in a sense be embodied (see, for example Wellner 2020, 3) This surely is a reasonable position, however, here the point is that for the AI-system to be conscious or even self-conscious it would need to be aware of its *own* boundaries.

In summary, at least the following two observations can be made from the above discussion. First, present AI systems are still basically constructed as input-output systems. Mitchell's work at least points in the direction that overcoming this single directedness via, for example, feedback processes could play an essential role in developing AGI systems. Closely related to this suggestion is the idea that embodiment and situatedness may play a central role in the development of AGI. However, the concepts of embodiment and situatedness lead to a more philosophically-oriented discussion of what may be necessary to host a mind.

1.3.1. Some results from research on the human mind

One of the central questions in the philosophy of mind is how the human mind arises or emerges from its underlying correlating structure. Often the philosophical discussion is framed in terms of the plausibility of various metaphysical positions such as physicalism, emergentism, panpsychism, dualism, idealism, and so forth.⁴ The underlying structure is, in the case of humans and animals, obviously the brain. Here I presuppose that some form of emergentism is the most plausible position. In the following, I present results from four researchers working with questions concerning the human mind and human consciousness. Their findings will be extrapolated to the question of which features are necessary for the realization of AGI.

1.4. Giulio Tononi's integrated information theory

Tononi introduced the Integrated Information Theory (ITT), a theoretical framework that focuses on the structure and the organization of the underlying substance of a conscious being. In his ITT, Tononi expresses that it is how much integrated information a system incorporates that matters for the emergence of consciousness. This integrated information is in turn dependent on the structure and organization of the system (for example, see: Tononi 2008; Tononi, Oizumi, and Albantakis 2014). Importantly, apart from the claim that the underlying substance is not crucial, but rather its structure, it is also argued in Tononi's theory that mere functionality is not sufficient for the creation of consciousness. A feed-forward system, that is, an input-output system, which in terms of functionality produces an adequate *output* to any given task, can be shown *not* to meet the requirements for consciousness within Tononi's theory (Tononi, Oizumi, and Albantakis 2014; Tononi and Koch 2015). Thus Tononi's theory would support the following claims concerning

AGI: (1) it should, at least in principle, be possible to create a thinking, sentient, and conscious being, or an AGI, not based on biological structures, (2) it is the structure of the system that matters, not its functionality; in particular, input-output systems would not be classified as conscious according to Tononi's ITT.

1.4.1. Thomas Metzinger's self-model

Philosopher of mind Thomas Metzinger made a detailed attempt to develop a theory of consciousness and also self-consciousness based on the *phenomenology* of consciousness. His starting point is an understanding of mental *representation*, which involves the notion of self-reflexivity.⁵ It should be noted here that, for example, Brooks, in the 90's, in the field of AI, proposed that AI need *not* be based on representations (Brooks 1991). However, it seems that what Brooks had in mind when using the term "representation" was abstract propositions about an object such as "(CAN (SIT-ON PERSON CHAIR)), (CAN (STAND-ON PERSON CHAIR))" for the abstract description of a chair. These, as I understand it, Brooks considers to be too narrow (Brooks 1991, 143). Others, using a broader understanding of representations, have argued that representations are at least necessary if the AI-system is intended to be conscious (Müller 2007). Anyway, central in Metzinger's definition is that any state in a system—for example, a brain or an AI-system—which represents a state in the world, is part of the system itself as a whole. Thus it can itself become the object of other, higher-order, representational processes. These processes could in turn be involved in the control of actions. Metzinger's self-model and definitions of mental self-presentation, or phenomenal self-presentation, for example, are all based on this understanding of representation, and they all incorporate a *recursive* element (Metzinger 2004, 42,42,87,90,282-288).

Such claims about the importance of recursiveness have also recently been made in the context of AGI. A group of researchers has claimed that "Reward is enough" to finally create AGI. Their idea is that the central mechanism is "feedback based on reward maximization" and they base their claims partly on the success of biological systems to create consciousness and mentality (Silver et al. 2021). The recursive element in Metzinger's theory is clearly compatible with the idea of reward maximization by feedback or reinforcement, proposed by these researchers, who work within the DeepMind project (Silver et al. 2021). Note that Tononi's theory also claims that mere feed-forward processing of information is not sufficient for the creation of consciousness.

Based on his self-model, Metzinger suggests that any AI-system intended to be conscious should incorporate an integrated and dynamical *model of the world*⁶ (Metzinger 2014, 279). He also describes embodiment in terms of his self-model: "In human beings, it is particularly interesting to note how the self-model simultaneously treats the target system 'as an object' (e.g., by using proprioceptive feedback in internally simulating ongoing bodily movements) and 'as a subject' (e.g. by emulating its own cognitive processing in a way that makes it available for conscious access). This is what 'embodiment' means [...]" (Metzinger 2004, 301). Metzinger furthermore stresses that the processes involved in creating a self-model are ongoing *dynamical* processes (Metzinger 2004, 322, 563–64). I deduce that mental and phenomenal representations in humans or in artificial systems therefore are not only self-reflexive but that the underlying processes are ongoing in time.

It can be said that Metzinger's position, although he focuses on *representation* (in a broad sense), at least implicitly incorporates both embodiment and situatedness. His view furthermore adds and emphasizes the significance of involving self-reflexive or recursive elements, possibly also in a broader sense. Furthermore, his approach is entirely naturalistic and, as such, may have advantages in a dialogue with AI researches

⁵ The term mental representation can very broadly understood as "... a mental object with semantic properties" (Pitt 2020).

⁶ "Darum braucht jede Maschine, die die Eigenschaft des bewussten Erlebens aufweist, ein integriertes und dynamisches Weltmodell" (Metzinger 2014, 279)

⁴ For a state of art overview see for example *The Routledge Handbook of Consciousness*. (Gennaro 2018)

who, presumably at least implicitly, often presume a naturalistic point of view.

1.4.2. Alva Noë's approach to perception

Another approach to understanding the *human* mind, and which places a stronger emphasis on embodiment, can be found in the work of Alva Noë. Noë often refers to studies of the human mind in the natural sciences, in particular, the cognitive sciences. He emphasizes activity in perception and introduces the concept of “enactive approach to perception” (Noë 2004, chap. 1). In his reasoning, he stands closer to the above research in AI that questions the central role of representations. He proposes that the role of representations in perception at least needs to be reconsidered (Noë 2004, 22–24). As an example, it has been observed that children who have been blind all their life, and who regain eyesight, initially have visual perception that sees unordered color-patches. The children could not, for example, associate the concept of an orb with the visual picture represented in their eyes, although from tactile experience they knew what an orb is (Sinha 2013, 48–55; Held et al. 2011, 551–53). In other words, precisely as Noë claims, the representation of what is “seen” in the retina of the individual does *not* constitute the *perception* of what is seen. Representation may be part of the process of perception but perception cannot entirely be understood or modeled by representations. The fact that the children in the above-named experiment eventually acquired the ability to perceive, for example, by sight, emphasizes the significance and importance of the context; that is, of embodiment and situatedness. In the case of developing AGI systems in which a mind may emerge, this could mean that the representation of any given perception, as pixels or strings of bits, should not be seen as central; rather how this representation interrelates to other representations seems to matter and to establish the actual meaning of the representation.

Another closely related example, which hints a further direction for the development of AGI, is the following. In the cognitive sciences, two well-known phenomena are that memories are often imprecise or even false, and that reasons for decisions or choices are often constructed in hindsight (for example, Wade et al. 2002; Oakes and Hyman 2001; Johansson et al. 2005; Hall et al. 2010). Like Noë's suggestion that perception involves activity in the person who perceives, these two phenomena can be interpreted in terms of constructive or creative activity in the processes of remembering or explaining decisions. Furthermore, in either case, the constructive or creative activity of remembering or in providing reasons, is aimed at creating an ordered and coherent mental image of the world in its totality. It seems, that the human mind, in the processes of perception, memorizing, providing reasons and, presumably, in mental processes in general, consciously or unconsciously, attempts to create a coherent and consistent inner image of the world, which correlates with the outer world. This process of creating an inner image of the world can, as I have argued in a different academic context, be regarded as a re-creation of the world (Li 2020). However, an inner image of the world is, since the self is part of the world, both partly an image of the self, and at the same time it is the self's image of the world. In parallel to Metzinger's reasoning about subject and object mentioned above, human consciousness of the world always has a part, which objectively is about the outside world and subjectively is about the self *in* the world. Noë similarly questions the strict distinction between what goes on inside us and outside us: “[...] one can reasonably wonder why we find it so plausible that there could be a consciousness like ours independent of the active exchange with the world? Why are we so certain consciousness depends only on what is going on inside us?” (Noë 2004, 211).

To sum up, the above discussion further emphasizes the significance of embodiment and situatedness for consciousness. Furthermore, it has been suggested that the creation of an inner image of the world, which correlates with the outer world, is significantly involved in perception processes.

1.4.3. Anil Seth's controlled hallucination

In his recent monograph *Being You*, neuroscientist Anil Seth has also proposed that consciousness, similar to the above creation of an inner world, can be understood as a kind of controlled hallucination (Seth 2021, chap. 4). Seth introduces the term “controlled hallucinations” based on ideas similar to those proposed here. However he does not emphasize the role and, I believe, importance of recurrency and feedback as strongly as I suggest it should be. Also, although Seth combines the term “hallucination” with “controlled”, the term hallucination also bears negative connotations and one could easily associate it with the idea that the inner world created by human consciousness is not real, less real, or somehow flawed. Still, even if the inner world is created — or, as I suggest, recreated by the mind — the correlation to the outer world is *real* and reflects relevant aspects of the outer world. If this were not the case, then how could we even *know* anything about the outer world?

Be that as it may, in summary, the all of the above examples highlight the following features of the *human* mind: embodiment, situatedness, self-reflexivity (in contrast to mere input-output systems), and creation or re-creation of an inner world. It is also clear that the human conscious system is *not* an input-output system. Also, it should be noted that *intelligence* is *not* included in these features. Indeed, it seems that the creation of mind does not *necessarily* go hand in hand with evermore advanced exhibitions of intelligent behavior. This has also briefly been discussed by Seth (Seth 2021, 247–51).

In relation to the above features, at least two further questions appear to be of significance. Firstly, are these features grounded in an overly anthropocentric view of what is needed to host a mind? If this is the case, then they may not be of greater relevance for the project of developing AGI. Secondly, how do these features relate to the question of creating an artificial mind? In the next section, I shall now turn to the first question.

1.5. Reflections on anthropocentrism in relation to embodiment, situatedness, self-reflexivity, and the re-creation of an inner world

Initially, it can be noted that the two features of self-reflexivity and the re-creation of an inner world relate to the following suggestion made above: that feedback processes might play an important role in the development of AGI systems. The features of embodiment and situatedness should, if they are of importance for crossing the barrier of meaning in AI, somehow be connected to the ability to understand, to make analogies, to form abstractions, and to other related cognitive abilities.

In the case of humans and, I assume, in animal intelligence alike, such a connection between meaning and understanding on the one hand, and embodiment and situatedness on the other hand, seems rather obvious. How could a human *understand* what a table is without an external relationship to a table, without placing the table in a *context*?⁷ How could you know what it is to be cold if you have not been exposed to cold in a body which can perceive coldness and in a real situation?⁸

⁷ This example has a parallel in anthropological research. Researchers in anthropology commonly face the problem that they may not properly understand certain aspects of the culture they are studying due to the *lack* of experiencing the culture from *within*.

⁸ In fact the famous “knowledge argument” for the qualitiveness of experience is based on this intuition. A scientist called Mary “knows” all (physical) facts about colors, but has never *experienced*, for example, the color red. But then her knowledge of all facts cannot be complete, for she does not know “what it is like” to see red; she does not know the actual “meaning” of red (Chalmers 2001, 7). The standard conclusion is that there is more to “red” than the mere physical facts. Here, I believe, the knowledge argument points to the thought that since there seems to be more to the meaning of red than the mere propositional representation, red needs to be experienced in a context, in a situation, via a body. On a more abstract and theoretical level, Tononi in his IIT also argued for the importance of interrelatedness of consciousness and qualitative experience (Tononi 2008, chaps. 238–239). Likewise, Metzinger has suggested that the

However, even if being in a body and factually experiencing, for example, coldness may be necessary for *humans* to grasp the meaning of redness, why should this also be the case in artificial intelligence? Isn't the proposal that embodiment and situatedness are *necessary* for the development of AGI an overly anthropocentric assumption?

In the above example of the GPT, one can pose the question of what would be needed for the machine to *understand* what love is. The GPT described love as “a feeling that people experience”. Surely, both feeling and experience seem to *presuppose* embodiment and situatedness. If the GPT would have experienced love in a body and in situations, then it would at least have the possibility to know what love actually *means*. Also, the GPT would presumably *not* produce an entirely different answer to a second request to define what love is if it were *situated*. Rather, it might answer more adequately something like: “Haven't I explained this just a few minutes ago?” or at least it would produce a similar answer. So embodiment and situatedness do not seem to be unreasonable requirements for an AGI.⁹

On a more abstract level independent from human life, embodiment and situatedness can be seen as *being-in space* and *being-in time*. Apparently, having a body entails being in space, having a spatial location, while being situated entails both being in space and being located in a temporal process. Space and time are, at least in modern physics, regarded as fundamental. Of course, in philosophy, it has also been argued that there is a fundamental connection of space and time to subjectivity and all appearances in our minds. One important example would be the explications by Kant in his *Critique of Pure Reason*. These explications on space and time strongly suggest that Kant thinks that a conscious mind needs to be able to relate, to live, to act in space and in time (Kant 1787 KrV B 41, 49-51).

This line of thinking is further supported by, for example, the existentialist views on consciousness by Jean-Paul Sartre and Martin Heidegger. Sartre points out the intimate connection between consciousness and being of the world in general (Sartre 2003 [1943], 17-18). In my reading, he suggests that consciousness – any consciousness – cannot exist without a being ‘other than itself’, that is, a world in which consciousness is embodied and situated. Similarly, Heidegger understands ‘to-be-in-the-world’ (In-der-Welt-sein) as the basic constitution of ‘Dasein,’ which I take to be his understanding of consciousness. In particular, Heidegger claims that there is no subject – in my reading, a first-person consciousness – without a world, that is, ‘being-in-the-world’ (Heidegger 1979 [1927] chap 2, 4 Section 12, 25).

If the above philosophers are correct, then embodiment and situatedness would not only be necessary for human minds but for *any* mind. Surely, one could argue that any form of reasoning, be it in physics, psychology or, for example, philosophy, obviously is based on our human preconditions and thus by default is anthropocentric. This objection, I believe, however, misses the point. Of course, as humans, we can *never* free ourselves from our human preconditions. However, space, time, and being-in-the-world may still be *universal* and universally necessary for subjectivity and appearances for whichever mind is in question, and the obvious way to relate to space would be via a situated body. To

observations made in so-called Ganzfeld experiments should be interpreted in terms of the importance of the *relatedness* of presentational content in perception (Metzinger 2004, 102-4).

⁹ It has recently been argued that even un-embodied and un-situated algorithms have some grounding: “[...] Google Translate and DeepL have no direct but an indirect grounding. They refer to the world indirectly” (Lyre 2020, 337). Lyre is not actually arguing against embodiment or situatedness. He merely points out that semantic grounding can be indirect. If this gives meaning in the context of indirect grounding, I believe that this meaning would be rather different than the meaning based on fully-grounded experience in real life. Furthermore, it is important to be aware of the reciprocal effects of AI-systems also on the *human* understanding of emotions, in this case love. As the postphenomenological approach emphasizes human relations to technology will also shape how we are as humans (see, for example, Liberati 2022)

be sure, the body of an AGI may be quite different from any biological body and even the time relation may be different. Processes in silico are generally orders of magnitude faster than neural connections, and an artificial body could, for example, be a space-station.

In particular, the situatedness in time would also entail a human or an artificial intelligence to be part of a life-story, a narrative and, since time extends to the future, being situated in time would also mean having some direction to the future. Humans perceive this directedness towards the future in terms of having imaginations and *intentions* for the future. These could be predictions, goals, or wishes, for example. Again, although having intentions seems to be based on human thinking, it still seems reasonable that, if an AGI has to be situated *in time*, then it has to have a relation both to the past and the future. Thus another feature would be the ability to have and create one's own intentions: *self-intentionality*. The above mentioned postphenomenological approach, as described by, for example Peter-Paul Verbeek, rather focuses on the possibility to move beyond a subject-object distinction. Intentionality, even human intentionality, would then be mediated by the technological artifact. A proposed AI-system would have a share of intentionality by the interaction of humans with machines (Verbeek 2011, 14-18, 55-58) However, the idea here is to create intentions, not foremost on the basis of the mediated intentions of humans, but more directly by the systems *own* intentions.

The feature of self-reflexivity seems to be independent of the system being human or not. Surely, feed-back processes occur and can be implemented in any kind of system, not merely biological systems. Thus, in this case, the objection of anthropocentrism does not seem to be applicable. Here the recent research by Silver et. al. provides an important perspective. They have argued that reward maximization provides a basis for understanding crucial features of general intelligence, and thus I take it also for AGI as, for example in knowledge learning, perception, social intelligence, imitation, or language (Silver et al. 2021). However, reward maximization is precisely a process based on feed-back and self-reflexivity.

The proposed re-creation of an inner world which correlates to and reflects the outer world, though, involves the notion of a world which again seems to rely on *being-in* a world, that is, on embodiment or situatedness. While one could raise a similar question about the anthropocentric focus of this feature, a similar response based on the universality of space and time can be given.

So it seems that there strong reasons to believe that embodiment, situatedness, self-reflexivity, the re-creation of an inner world, being part of a narrative, and self-intentionality—the latter could be seen as a special case of self-reflexivity and is also implied by embodiment and situatedness — are all necessary features not only for a human mind but also for an artificial mind. These features *appeared* to be strongly oriented towards the human mind. However, despite the fact that we as humans take our starting-point in human life, reasons have been proposed as to why these features are universal. Also, although these features may be *necessary*, they may not be sufficient for the creation of an artificial mind. So how do these features relate to the question of creating an artificial mind? Are there any hints about how these features can be implemented in artificial systems?

1.6. The features in an artificial system

In this Section I will turn to the suggested features and hints on how to implement them in an artificial system with the possible goal to create an artificial mind. In particular, I shall focus upon the proposed ‘re-creation of an inner world’. The implementation of the features of embodiment, situatedness, self-reflexivity, and self-intentionality all have in more or less developed forms been implemented or suggested even in relation to artificial systems. Embodiment and situatedness have, at least in a rudimentary form, been implemented in robotics. Lake et al., for example, stress that, among other features, an AGI would need to have “intuitive physics”, “intu-

itive psychology”, a sense for causality, and the ability to learn how to learn (Lake et al. 2017). The feature of self-reflexivity is very general and has been specified in greater detail by, for example, Lyre. He suggests six different dimensions of self-reflexivity: self-learning, self-repairing, self-replicating, self-exploratory, self-explanatory, and self-conscious (Lyre 2020, 342). Metzinger remarks that “[...] artificial systems as known today [2004] do not possess genuinely embodied goal representations [...]” (Metzinger 2004, 199 my comment).

Far more interesting is the proposed re-creation of an inner world that reflects the outer world. This surely is indirectly suggested in Metzinger’s theory of consciousness. Metzinger describes dreams, hallucinations, and stable “fully-awake” consciousness and self-consciousness in terms of being, in principle, the same processes. They all are based on “the appearance of a world” (Metzinger 2014, 38, 77-78, 206). Indeed, applied to the features which are necessary for the development of an AGI, his ideas about consciousness suggest that, among embodiment, situatedness, various forms of self-reflexivity including self-intentionality, being part of narrative, and the re-creation of a world internal to the system, the latter would play a central role.

If, as it seems, the re-creation of an inner world that reflects the outer world is crucial to creating AGI, the question arises as to how this re-creation can be implemented in present AI-structures. A possible answer to this question links back to Mitchell’s observation about feedback from higher layers in the visual cortex. Recall that the function of these feedback processes was not well understood. Interestingly, something, which at least loosely resembles these feedback processes, has been implemented in what is called Deep Dreaming (Hayes 2015; Mordvintsev, Olah, and Tyka 2015). The process can briefly be described as follows:

One way to visualize what goes on is to turn the network upside down and ask it to enhance an input image in such a way as to elicit a particular interpretation. Say you want to know what sort of image would result in “Banana.” Start with an image full of random noise, then gradually tweak the image towards what the neural net considers a banana (Mordvintsev, Olah, and Tyka 2015).

or

Here’s a recipe for deep dreaming. Start by choosing a source image and a target layer within the neural network. Present the image to the network’s input layer, and allow the recognition process to proceed normally until it reaches the target layer. Then, starting at the target layer, apply the back-propagation algorithm that corrects errors during the training process. However, instead of adjusting connection weights to improve the accuracy of the network’s response, adjust the *source image* to increase the amplitude of the response in the target layer (Hayes 2015 my emphasis).

This method produces images which are made up of a “psychedelic” mixture of the categories the network is trained to identify and the original input. In a sense, it allows the human to “see” what is going on in the artificial neural network while processing a picture (Mahendran and Vedaldi 2015; Mordvintsev, Olah, and Tyka 2015). Hayes prefers to call these psychedelic images produced by this process hallucinations:

(H)ere the visual system is hyperactive, and what it generates are hallucinations. In these images we witness a neural network struggling to make sense of the world. The training process has implanted expectations about how pieces of reality should fit together, and the network fills in the blanks accordingly (Hayes 2015).

However, if, as Hayes seems to suggest, the network has “expectations about how pieces of reality should fit together”, then these expectations could possibly be used in a more systematic manner. Presumably, they could be used to achieve the re-creation of a world internal to the system. To better see this, imagine the following: My inner picture of the world, for example, after waking up in the morning is not merely a result of the actual sense perceptions I have, but also the result of my

expectations of what I might perceive. If I wake up in a foreign place, I may initially expect to see my bedroom at home and for a short moment I might be disoriented until I have adapted to the novel situation. In other words, my expectations *meet* the input, and if the input and the expectations do not correlate the latter are modified.

One difference between this scenario and artificial systems such as the Deep Dreaming network is surely that both the Deep Dreaming network and standard artificial neural networks, although they involve recurrent processes, still essentially are input/output based. The Deep Dream algorithm feeds back a proposed output into the source image. For example, if the algorithm is expected to “recognize” cats then cats will appear in the source image in unexpected places; the source image is modified and appears “psychedelic”. However, in a re-creation, the feed-back would be based on a broad variety of expectations. The source image, which would have been modified by the feed-back in general, would correlate and would be *supposed to correlate* with the original input, thus re-creating a consistent inner image of the world. The inner image, although recreated, would fit the input in accordance to previous expectations. The process would, and I believe *should*, be dynamical and inherently recursive. The Deep Dreaming network, however, still produces an output to a given input: the psychedelic pictures mentioned above.¹⁰

Thus, I believe, that the idea in the Deep Dreaming network of feeding data “backwards” could possibly be modified. How this might be implemented, and if this is successful, is, of course, a matter for empirical investigation and requires the skills of advanced programming. In any case, independent of whether an approach based on back-propagation and Deep Dreaming, possibly with multiple “outputs” back-propagated into the alleged inner image of the outer world, turns out to be a possible way of creating an inner world in a machine, I believe that the feature of re-creating the world in an inner world is an important feature in the creation of a mind. Indeed, re-creating the world in an inner-world, if it is implemented in a machine, would turn the machine into a device that is not merely an input/output device in the traditional sense, which – recall the conclusion from Tononi’s IIT – cannot be conscious, but into a device which dynamically connects the input with the output and thus possibly could be conscious. The dynamic connection can again be seen as a form of self-reflexivity. It seems that even in the case of recreating the world as an inner world there at least are hints in current technology to how or in which direction one may proceed.

2. Conclusion

In summary, I have argued that at least the following features need to be incorporated in a system if it is to become conscious: embodiment, situatedness, being part of a narrative, self-intentionality, self-reflexivity, and the re-creation of an inner world. Being part of a narrative, and self-intentionality, express the time-aspect of embodiment and situatedness. Since, as I believe, space and time should be regarded as *universal* aspects of existence not based on an overly anthropocentric view, I also believe that the temporal aspects of embodiment and situatedness, in being part of a narrative and having goal-directed self-intentionality, are worth emphasizing. Importantly, in all the discussions, be it the more philosophical or the more technical oriented parts, self-reflexivity has been strongly emphasized. Self-reflexivity is a reoccurring theme. The re-creation of the world as an inner world seems to be impossible without involving some form of feedback mechanism. Even on a theoretical level, as in Tononi’s IIT, simple feed-forward process do not suffice for creating consciousness.

¹⁰ A more recent technological approach attempts to bind together data from images, text, audio, depth, thermal, and IMU data (Girdhar et al. 2023). Such attempts could certainly be of relevance in the ‘re-creating the world in an inner world’, as proposed here.

In the case of the re-creation of the world as an inner world, I have hinted that the Deep Dreaming approach in programming might be worth developing in the direction of creating an inner image of the world. Indeed, without an inner image, which presumably would include an image of the system itself, it seems that “nobody would be at home”. The ideas about consciousness, both philosophical and more technical, clearly point in the direction of the importance of the creation of an inner image of the world, or a self-model as Metzinger expresses it. Thus, I also believe that while being embodied and situated as being part of the world are crucial in some sense, the *re-creation* of the outer world in an inner image based on *self-reflexive* processes is the most central part in any project of developing an AGI. Programs like Open AI’s GPT, impressive though they may be, are still constructed as input/output machines and lack an inner image of the outer-world.¹¹

To be sure, skeptics may nevertheless argue that the entire project of AGI is futile; after all, a machine is *merely* a machine. However, this seemingly simple objection only holds on the premise that one can accept the counter-intuitive conclusion that consciousness, self-consciousness and related mental states do not have an ontological status of their own. For if one believes that a mind can *never* emerge in a machine, merely because it is “a pack of integrated circuits” (paraphrasing Francis Crick’s famous quote that we “[...] are nothing but a pack of neurons” (Crick 1994), then one might as well believe that a mind can *never* emerge in a *human* since the brain is merely “a pack of neurons”. Or one would have to show that the *human* mind is unique in the sense that no other substrate than a biological brain has the quality to let a mind emerge, and that surely is overly anthropocentric.

Nevertheless, I do believe that a warning is also required. As Metzinger has repeatedly suggested, humans should be cautious in attempting to create AGI artificial minds, artificial consciousness, or a synthetic phenomenology for ethical reasons (Metzinger 2017; 2021). He writes: “On ethical grounds, we should not risk a second explosion of consciousness suffering on this planet, at the very least not before we have a much deeper scientific and philosophical understanding of what both consciousness and suffering really are” (Metzinger 2021, 46). We simply do not really know what kinds of suffering we might produce in the process of pursuing AGI. Which path do we wish to choose? We might choose the path of identifying the necessary conditions for artificial consciousness and eventually, perhaps, the sufficient conditions, and implementing them. Or we might choose the route of identifying the necessary conditions and, with the knowledge of possible problems, refrain from implementing them, thus retaining our human dignity in relation to creation and what we wish to create. I believe the latter is the path to follow.

Availability of data and material (data transparency)

Not applicable

Code availability

Not applicable

¹¹ According to a recent article, AI researcher Yann LeCun has proposed a ‘new vision’ for the development of future AI. LeCun uses the term ‘world model’ and compares it to animal brains: “LeCun thinks that animal brains run a kind of simulation of the world, which he calls a world model” (Heikkilä & Heaven, 2022). I take it that the research suggested in this article is in the same direction as the ideas of LeCun. However, I also believe that one should not talk of simulations; recreating an inner world as a reflection of the outer world is not merely a question of simulating something. The inner worlds humans or other biological beings re-create are real in a deep and profound sense.

Authors’ contributions

Not applicable

Declaration of Competing Interest

The authors declare that there are no conflicts of interest.

Acknowledgments

The research in this paper is funded by the Wallenberg Foundations WASP-HS program within the projects “Artificial Intelligence, Democracy and Human Dignity” and “The Artificial Public Servant”.

References

- Abramson, Darren. 2011. “Philosophy of Mind Is (in Part) Philosophy of Computer Science”. *Minds and Machines* 21 (2): 203–219. doi:10.1007/s11023-011-9236-0.
- Altman, Sam. 2023. “Planning for AGI and Beyond”. Open AI (blog). March 24 2023 <https://openai.com/blog/planning-for-agi-and-beyond>.
- Alpaydin, Ethem. 2016. *Machine Learning: The MIT Press*.
- Ayfer, Murat. 2020. “Philosopher AI.” 2020. <https://philosopherai.com/>.
- Baldassarre, Gianluca, Giuliano Santucci, Vieri, Cartoni, Emilio, Caligiore, Daniele. 2017. “The Architecture Challenge: Future Artificial-Intelligence Systems Will Require Sophisticated Architectures, and Knowledge of the Brain Might Guide Their Construction”. *Behavioral and Brain Sciences* 40: e254. doi:10.1017/S0140525X17000036.
- Beniagiev, David, Segev, Idan, London, Michael. 2021. “Single Cortical Neurons as Deep Artificial Neural Networks”. *Neuron* 109 (17): 2727–2739. doi:10.1016/J.NEURON.2021.07.002, e3.
- Bosselut, Antoine, Rashkin, Hannah, Sap, Maarten, Malaviya, Chaitanya, Celikyilmaz, Asli, Choi, Yejin. 2020. “CoMET: Commonsense Transformers for Automatic Knowledge Graph Construction”. In: *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, 4762–4779*. doi:10.18653/v1/p19-1470.
- Bostrom, Nick. 2014. *Superintelligence: Oxford University Press*.
- Brooks, Rodney A. 1991. “Intelligence without Representation”. *Artificial Intelligence* 47 (1): 139–159. doi:10.1016/0004-3702(91)90053-M.
- Brown, Tom B, Mann, Benjamin, Ryder, Nick, Subbiah, Melanie, Kaplan, Jared, Dhariwal, Prafulla, Neelakantan, Arvind, et al.. 2020. “Language Models are Few-Shot Learners”. *Advances in Neural Information Processing Systems 2020-Decem*.
- Chalmers, David J. 1996. *The Conscious Mind: Oxford University Press*.
- Chalmers, David J. 2001. “Consciousness and its Place in Nature”. *Cognitive Systems Research* 2 (2): 177. doi:10.1016/S1389-0417(01)00028-6.
- Charniak, Eugen. 2018. *Introduction to Deep Learning: The MIT Press*.
- Coeckelbergh, Mark. 2020. *AI Ethics: The MIT Press*.
- Crick, Francis. 1994. *The Astonishing Hypothesis, New York: Charles Scribner’s Sons*.
- Crosby, Matthew. 2020. “Building Thinking Machines by Solving Animal Cognition Tasks”. *Minds and Machines* doi:10.1007/s11023-020-09535-6.
- Floridi, Luciano, Chiriatti, Massimo. 2020. “GPT-3: Its Nature, Scope, Limits, and Consequences”. *Minds and Machines* 30 (4): 681–694. doi:10.1007/s11023-020-09548-1.
- Gennaro, Rocco J. 2018. *The Routledge Handbook of Consciousness: Routledge Edited by Rocco J. Gennaro*.
- Girdhar, Rohit, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. “ImageBind: One Embedding Space To Bind Them All.” arXiv.
- Hall, Lars, Johansson, Petter, Tärning, Betty, Sikström, Sverker, Deutgen, Thérèse. 2010. “Magic at the Marketplace: Choice Blindness for the Taste of Jam and the Smell of Tea”. *Cognition* 117 (1): 54–61. doi:10.1016/j.cognition.2010.06.010.
- Hayes, Brian. 2015. “Computer Vision and Computer Hallucinations”. *The American Scientist* 103 (6). doi:10.1038/194733b0.
- Heidegger, Martin. 1979. *Sein Und Zeit [1927], Tübingen: Verlag Max Niemeyer*.
- Heikkilä, Melissa, Heaven, Will Douglas. 2022. “Yann LeCun Has a Bold New Vision for the Future of AI”. *MIT Technology Review* 2022. <https://www.technologyreview.com/2022/06/24/1054817/yann-lecun-bold-new-vision-future-ai-deep-learning-meta/>. accessed 2022-06-27.
- Held, Richard, Ostrovsky, Yuri, DeGelder, Beatrice, Gandhi, Tapan, Ganesh, Suma, Mathur, Umang, Sinha, Pawan. 2011. “The Newly Sighted Fail to Match Seen with Felt”. *Nature Neuroscience* 14 (5): 551–553. doi:10.1038/nn.2795.
- Johansson, Petter, Hall, Lars, Sikström, Sverker, Olsson, Andreas. 2005. “Failure to Detect Mismatches between Intention and Outcome in a Simple Decision Task”. *Science (New York, N.Y.)* 310 (5745): 116–119. doi:10.1126/science.1111709.
- Kant, Immanuel. 1787. *Kritik der reinen Vernunft*.
- Kaplan, Andreas, Haenlein, Michael. 2019. “Siri, Siri, in My Hand: Who’s the Fairest in the Land? On the Interpretations, Illustrations, and Implications of Artificial Intelligence”. *Business Horizons* 62 (1): 15–25. doi:10.1016/j.bushor.2018.08.004.
- Lake, Brenden M, Ullman, Tomer D, Tenenbaum, Joshua B, Gershman, Samuel J. 2017. “Building Machines That Learn and Think like People”. *Behavioral and Brain Sciences* 40: E253. doi:10.1017/S0140525X16001837.

- Lenzen, Manuela. 2018. *Künstliche Intelligenz*: C.H.Beck.
- Liberati, Nicola. 2022. "Digital Intimacy in China and Japan". *Human Studies* doi:10.1007/s10746-022-09631-9.
- Li, O. 2020. "Creativity in Process-Panpsychist Panentheism, and the Mind". *Philosophy, Theology and the Sciences* 7 (1): 30–47. doi:10.1628/ptsc-2020-0004.
- Lyre, Holger. 2020. "The State Space of Artificial Intelligence". *Minds and Machines* 30 (3): 325–347. doi:10.1007/s11023-020-09538-3.
- MacLennan, Bruce James. 2017. "Benefits of Embodiment". *Behavioral and Brain Sciences* 40: e271. doi:10.1017/S0140525X17000206.
- Mahendran, Aravindh, Vedaldi, Andrea. 2015. "Understanding Deep Image Representations by Inverting Them". In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 07-12-June doi:10.1109/CVPR.2015.7299155, 5188–96.
- Metzinger, Thomas. 2004. *Being No One*: The MIT Press.
- Metzinger, Thomas. 2014. *Der Ego-Tunnel*: Piper Verlag.
- Metzinger, Thomas. 2017. "Benevolent Artificial Anti-Natalism (BAAN)." Edge. 2017. https://www.edge.org/conversation/thomas_metzinger-benevolent-artificial-anti-natalism-baan.
- Metzinger, Thomas. 2021. "Artificial Suffering: An Argument for a Global Moratorium on Synthetic Phenomenology". *Journal of Artificial Intelligence and Consciousness* 08 (01): 43–66. doi:10.1142/s270507852150003x.
- Mitchell, Melanie. 1993. *Analogy-Making as Perception*: The MIT Press.
- Mitchell, Melanie. 2019. *Artificial Intelligence*: Straus and Giroux Farrar.
- Mordvintsev, Alexander, Christopher Olah, and Mike Tyka. 2015. "Inceptionism: Going Deeper into Neural Networks." The Googel AI Blog. 2015. <https://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html>.
- Müller, Vincent C. 2007. "Is There a Future for AI without Representation?" *Minds and Machines* 17 (1): 101–115. doi:10.1007/s11023-007-9067-1.
- Mykhailov, Dmytro, Liberati, Nicola. 2022. "A Study of Technological Intentionality in C++ and Generative Adversarial Model: Phenomenological and Postphenomenological Perspectives". *Foundations of Science* doi:10.1007/s10699-022-09833-5.
- Noë, Alva. 2004. *Action in Perception*: The MIT Press.
- Oakes, Mark A, Hyman Jr, Ira E. 2001. "The Role of the Self in False Memory Creation". *Journal of Aggression, Maltreatment & Trauma* 4 (2): 87–103. doi:10.1300/J146v04n02_05.
- Pasquinelli, Matteo. 2019. "How a Machine Learns and Fails: A Grammar of Error for Artificial Intelligence". *Spheres. Journal for Digital Cultures* 5: 1–17 Spectres of AI.
- Pitt, David. 2020. "Mental Representation". In: Edward, N Zalta (Ed.), *The {Stanford} Encyclopedia of Philosophy*: Metaphysics Research Lab, Stanford University edited by(S)pring 2.
- Sap, Maarten, Le Bras, Ronan, Allaway, Emily, Bhagavatula, Chandra, Lourie, Nicholas, Rashkin, Hannah, Roof, Brendan, Smith, Noah A, Choi, Yejin. 2019. "ATOMIC: An Atlas of Machine Commonsense for If-Then Reasoning". In: *Association for the Advancement of Artificial Intelligence*, 3027–3035. doi:10.1609/aaai.v33i01.33013027.
- Sartre, Jean-Paul. 2003. *Being and Nothingness* [1943]: Routledge.
- Searle, John R. 2004. *Mind: A Brief Introduction*: Oxford University Press.
- Seth, Anil. 2021. *Being You*: Faber.
- Silver, David, Singh, Satinder, Precup, Doina, Sutton, Richard S. 2021. "Reward Is Enough". *Artificial Intelligence* 299, 103535. doi:10.1016/j.artint.2021.103535.
- Singer, Peter. 2009. *Animal Liberation. An Imprint of*: Harper Collins Publishers.
- Sinha, Pawan. 2013. "Once Blind and Now They See". *Scientific American Magazine* 309: 48–55. doi:10.1038/scientificamerican0713-48, July.
- Tononi, Giulio. 2008. "Consciousness as Integrated Information: A Provisional Manifesto". *Biological Bulletin* 215 (3): 216–242. doi:10.1038/215/3/216.
- Tononi, Giulio, Koch, Christof. 2015. "Consciousness: Here, There, and Everywhere". *Philosophical Transactions Biological Sciences* 370 (1668). doi:10.1098/rstb.2014.0167.
- Tononi, Giulio, Oizumi, Masafumi, Albantakis, Larissa. 2014. "From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0". *PLoS Computational Biology* 10 (5). doi:10.1371/journal.pcbi.1003588.
- Verbeek, Peter-Paul. 2011. *Moralizing Technology*, Chicago: The University of Chicago Press.
- Wade, Kimberley a, Garry, Maryanne, Don Read, J, Stephen Lindsay, D. 2002. "A Picture Is Worth a Thousand Lies: Using False Photographs to Create False Childhood Memories". *Psychonomic Bulletin & Review* 9 (3): 597–603. doi:10.3758/BF03196318.
- Watson, David. 2019. "The Rhetoric and Reality of Anthropomorphism in Artificial Intelligence". *Minds and Machines* 29 (3): 417–440. doi:10.1007/s11023-019-09506-6.
- Wellner, Galit. 2020. "Material Hermeneutic of Digital Technologies in the Age of AI". *AI & SOCIETY* doi:10.1007/s00146-020-00952-w.
- Wilson, Robert A, Foglia, Lucia. 2017. "Embodied Cognition". In: Edward, N Zalta (Ed.), *The {Stanford} Encyclopedia of Philosophy*: Metaphysics Research Lab, Stanford University edited by Spring 201.