Theoretical Philosophy
1 Minds, Brains, and Deep Learning: The Development of Cognitive Science Through the Lens of Kant’s Approach to Cognition

Abstract: This paper reviews several ways in which Kant’s approach to cognition has been influential and relevant for the development of various paradigms in cognitive science, such as functionalism, enactivism, and the predictive processing model of the mind. In the second part, it discusses philosophical issues arising from recent developments in artificial intelligence in relation to Kant’s conception of cognition and understanding. More precisely, it investigates questions about perception, cognition, learning, understanding, and about the age-old debate between empiricists and rationalists in the context of so-called deep neural network architectures as well as the relevance of Kant’s conception of cognition and understanding for these issues.

1 Introduction

If you follow the headlines, you can easily get the impression that much of contemporary cognitive science is heavily influenced by Kant’s philosophy. Philosopher Andrew Brook (1994) called him the “intellectual godfather” of cognitive science, since Kant allegedly already defended a functionalist theory of mind, arguably the philosophical foundation of artificial intelligence. Neuroscientist Georg Northoff (2018, viii) reports that rereading Kant’s Critique of Pure Reason has awakened him from his dogmatic slumbers, just like reading Hume had awakened Kant. Impressed by empirical evidence about self-generated brain activity, Northoff and others speak of the “Kantian brain” and associate this activity with Kant’s notion of spontaneity (Fazelpour/Thompson 2015). Francisco Varela (Weber/Varela 2002) acknowledged Kant’s enormous influence on his own autopoietic approach to life and cognition, and more recently Link Swanson (2016) has traced the popular predictive processing paradigm back to Kant’s general project. This is striking, given that Kant’s project was not primarily concerned with issues in the philosophy of mind but driven rather by epistemological
concerns. But although Kant may not have subscribed to all these views attributed to him, such writings present various ideas from his theoretical philosophy as having had or still having an enormous influence on contemporary philosophy of mind and cognitive science.

In this review paper, I will first sketch several ways in which Kant’s approach to cognition has been influential and relevant for the development of cognitive science. Kant’s relevance goes well beyond some vapid and superficial similarity of certain concepts; many philosophers claim that Kant already anticipated several tenets of classical cognitivism, enactivism, and the predictive processing model of the mind. In the second part, I will add one more piece to this story by discussing philosophical issues arising from recent developments in artificial intelligence. More precisely, I want to sketch some of the philosophical issues associated with so-called deep neural network architectures and the relevance of Kant’s conception of cognition and understanding for these issues. As will become clear, the performance of deep neural networks (DNNs) raises important questions about perception, cognition, learning, understanding, and about the age-old debate between empiricists and rationalists; this has led some researchers in machine learning to revive some of Kant’s core ideas regarding cognition, developing a Kantian cognitive architecture to overcome the shortcomings of existing deep learning architectures.

2 Cognitive Science Through the Lens of Kant’s Theoretical Philosophy

Kant’s general influence on contemporary thinking is unquestioned and familiar. Gomes (2017) lists an impressive number of mental phenomena for which Kant’s philosophy has been and still is very influential, e.g., the connection between consciousness and self-consciousness (Schlicht 2016/2017) or the debate about conceptual and non-conceptual perceptual content (McDowell 1994, Hanna 2008). Brook (1994) already considered several of Kant’s central claims.

Moreover, Gomes (2017) emphasizes the strong influence on specific philosophers in the 20th century, like Strawson and Sellars, who have then shaped the development of analytic philosophy. Most recently, a special issue of Synthese (198, Suppl. 13, 2021) brings together several authors discussing the relevance of various Kantian ideas, most notably his method of transcendental argument, for issues in the metaphysics of grounding, the use of the imagination and modal knowledge, virtue epistemology, ethics, and others.
about the mind as having fueled cognitive science more directly; most notably the claim that “most representations require concepts as well as percepts”, and Kant’s method of transcendental argument, understood as the attempt to “reveal the conditions necessary for some phenomenon to occur” (Brook 1994, p. 12). Based on this initial familiarity of Kant’s stance on issues in the philosophy of mind, cognitive science and contemporary debates, one can reconstruct the historical changes that cognitive science underwent through the lens of various aspects of Kant’s theoretical philosophy and find traces of some specific ideas of his thinking in the works of cognitive scientists.

2.1 Kant and Functionalism

When John McCarthy coined the term ‘artificial intelligence’ (AI) in the context of the famous Dartmouth conference in 1956, he described the goal of this project as “that of making a machine behave in ways that would be called intelligent if a human were so behaving” (McCarthy et al. 1955). In a similar vein, Margaret Boden describes the overarching goal of research in AI as “to make computers do the sorts of things that minds can do” (Boden 2016, p. 1). The focus in the first research phase that followed was already set by McCarthy et al. They intended to “attempt [. . .] to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves” (McCarthy et al. 1955, p. 12). While this has been achieved in some areas like speech production and chess computers, in which AI systems sometimes even outperform humans in very specialized problem-solving tasks, the “holy grail” (Boden 2016, p. 18) of AI research has always been the development of an AI system that exhibits “general intelligence”, understood “as the ability to perform tasks and attain goals in a wide variety of environments” (Shanahan 2019, p. 91, cf. Legg/Hutter 2007). This broad-stroke characterization of intelligence bypasses the apparent vagueness of the notion which may otherwise yield “terminological quibbles” (Walmsley 2012, p. 3). Walmsley does not regard the terminological choice of “intelligence” as significant but thinks that – echoing McCarthy’s goal – “the central issue of AI [. . .] is a comparative one: whatever we (humans) have, whether we call it intelligence’, ‘thinking’, ‘cognition’, ‘mind’, or something else, can machines have it too?” (Walmsley 2012, p. 3).2

2 The notion of intelligence is of course vague and difficult to operationalize, let alone compare across very different kinds of philosophical systems. Kant’s notion of (an) intelligence is not very helpful for the present project, since calling a creature “intelligent”, in his view, amounts to
Boden’s and Walmsley’s characterizations put computers and machines into focus. But, as Boden observes, computers or machines themselves aren’t what matters. AI is not about hardware, but about what artificial hardware can do. Therefore, the focus is not on machines, but on virtual machines (Boden 2016, p. 3), which are nothing but information-processing systems that can be implemented in a variety of hardware. Consequently, the favored philosophical background theory supporting the possibility of AI has been functionalism, according to which mental states in general are conceived in terms of their functions (or causal roles). Every mental state is identified by its set of causal relations to system inputs and outputs as well as other system states (Putnam 1965). The realization of this causal network of functions is taken to be contingent because the functions are considered to be multiply realizable (Polger/Shapiro 2016). Thus, Classical Cognitivism, the first paradigm in cognitive science, conceived of cognition as information processing along the lines of that present in digital computers. In particular, cognition was understood as constituted by syntactically driven manipulations of symbolic representational structures in the brain that are “sandwiched” (Hurley 1998) between sensory inputs and motor outputs (Fodor 1975, Pylyshyn 1984). For example, when I look at the coffee mug in front of me, sensory information hitting my retina is processed in specialized modules that eventually produce a detailed three-dimensional image of the mug that can guide actions like grasping it. Marr (1982) has provided an exemplary theory of perception in this regard.

One claim relevant for thinking about artificial intelligence is that Kant allegedly defended a functionalist conception of the mind. In Kant and the Mind, Andrew Brook (1994) interprets Kant’s agnosticism about the underlying substrate of the mind in this way. Despite Kant’s “implacably hostile” attitude towards materialism, Brook argues that “materialism fits remarkably easily into his overall theory” (Brook 1994, p. 15). Impressed by Kant’s position that “so far as the real nature of the mind is concerned, strict neutrality has to be the order of the day”, Brook takes this agnosticism to be an instance of the contemporary functionalist idea of the “multiple realizability” of mental functions; as do Sellars (1974) and Meerbote (1989).

However, this functionalist interpretation of Kant’s philosophy of mind faces some problems: Firstly, it ignores Kant’s peculiar conception of matter as mere appearance, which leads Ameriks (2000) to interpret Kant’s position as a considering it free and capable of self-determination, i.e., as belonging to the world of noumena rather than phenomena (GMS, p. 452, p. 458). Therefore, I suggest moving from intelligent to cognitive capacities for the purposes of this paper.
form of “mere immaterialism”. He arrives at this interpretation by allowing for a minimal knowledge about substrates of appearances, i.e., that they are not material. But this presupposition of knowledge of things in themselves, strictly rejected by Kant, makes Ameriks’ interpretation itself problematic. Secondly and relatedly, the functionalist interpretation of Kant’s philosophy of mind collides with the fact that contemporary functionalism is typically formulated as entirely ontologically neutral, but rather put forward as a stepping stone to materialist reductionism, since the analysis of mental phenomena in terms of their causal roles is usually complemented by an additional claim about (possibly multiple) physical realizations of these mental functions (Chalmers 1996/Kim 1998/Levine 2001/Block 2015). Interpreted this way, Kant would clearly oppose functionalism. Thirdly, it is questionable whether Kant would have taken all “functions” of the mind (KrV: A78f/B103f) to be ‘functionalizable’ in the sense required for being “realized” by a physical mechanism. For example, what Kant calls the “spontaneity” of mind, properly understood and characteristic of the understanding, seems incompatible with materialism. Allison, for example, is less optimistic than Brook and argues that Brook’s functionalist-materialist interpretation of Kant’s theory of mind cannot be right, since, in Kant’s view (or, rather Allison’s interpretation of it), “cognition must be conceived as more than an elaborate information processing procedure, one which begins with raw sensible input and ends with the relatively reliable products of the understanding (cognitions). [. . .] What is missing in such a picture of cognition (at least from the Kantian perspective) is precisely its self-conscious, apperceptive character” (Allison 1996, p. 63).

Whether there is a way of incorporating the notion of spontaneity (with or without its alleged intrinsic self-conscious aspect) and the unity of apperception within a broadly naturalist framework, is an interesting further question that I cannot pursue in depth here. Hanna and Thompson (2003), Northoff (2012), and Fazelpour and Thompson (2015) consider the brain’s self-generated activity as a candidate for a neural correlate of the function that Kant calls spontaneity, but this interpretation has not been justified in any detail (for a critical discussion see Schlicht & Newen 2015, cf. Northoff 2013/2014 for further connections to neuroscience).  

3 Indeed, ‘spontaneity’ is characteristic of ‘freedom’, which is typically contrasted to ‘nature’ and ‘mechanism’ in Kant’s works. Regarding the nature of the understanding, there is a debate about its absolute or merely relative spontaneity which raises the issue of whether cognition in general or acts of understanding in particular can be conceived as being caused by prior events. Grüne (2013) discusses these issues further.

4 Given such problems with a genuinely functionalist interpretation of Kant’s stance on mental phenomena, it is not surprising that a number of alternative interpretations of Kant’s view
2.2 Kant and the Cartesian Theatre in the Brain

Whether functionalism provides us with an accurate portrayal of the mind depends partly on the features of the biological implementation of mental functions in human (and animal) brains. Can cognition be conceived of as a set of causal functions in abstraction of the biological features of its realization, such that this set of functions could in principle be realized by a machine using a non-biological realization? Or is cognition a biological phenomenon whose realization depends on the presence of a complex dynamical biological system, namely, an organism (with a brain and nervous system), exhibiting crucially biochemical means of information processing? For example, might mental representational states be “aspects” of neural computations, i.e., biological, rather than being abstract functions enjoying some independence from their realizers (Piccinini 2020)?

In the 1980s, new imaging techniques in neuroscience initiated a research focus on the brain, resulting in connectionist neural network models of cognitive phenomena. They still remained computational and representational, but information (about the coffee mug in front of me, say) was now supposed to be processed subsymbolically; representations were proposed to have a non-linguistic structure (Smolensky 1988, Clark 1991, Sejnowski 1992, Churchland 1997). This turn was accompanied by new developments in robotics and artificial intelligence, since some researchers now rejected the need for full-fledged models of the world in favor of much sparser “subsumption architectures” that do not rely on a detailed representation of the world (Brooks 1991). As we will see in the second part of the paper, this turn towards brain architecture also inspired the more recent machine-learning techniques, with deep learning being the most prominent one.

Against the background of this controversy over functionalist and biological approaches to cognition, it is striking that on the one hand, Kant anticipated certain problems with the precursor to functionalism, namely the identity theory of mind and brain (Place 1956/Smart 1959), which later resurfaced as Dennett’s
‘Cartesian Theatre’ objection against materialism about consciousness, while on the other hand Kant was also impressed by the brain’s features that might explain certain cognitive phenomena. This tension can be brought to light by having a look at his exchange with the physician Samuel Thomas von Sömmerring.

In 1796, Samuel Sömmerring published a short book, *On the Organ of the Soul*, in which he speculated about the possible function of the liquid contained in the brain’s ventricles with respect to the unification and separation (synthesis and analysis) of sensory data. Prior to publication, he had an exchange with Kant about his ideas, specifically that of a sensory organ or seat of the soul in the brain. In one of his letters, Kant respects Sömmerring’s position but expresses his explicit doubts about the general approach, since “it is the concept of a *seat of the soul* that occasions the disagreement of the faculties concerning the common sensory organ and this concept therefore had better be left entirely out of the picture, which is all the more justified since the concept of a seat of the soul requires *local presence*” (AA 12, 31–32). In contrast to this approach, Kant suggests taking seriously the idea of a mere “*virtual presence*” of the mind in the brain, which makes the whole question of what could serve as a ‘seat of the soul’ disappear, or so he claims. Sadly, he does not clarify what he means by virtual presence here. A further striking passage in this regard can be found in his *Lectures on Metaphysics* (V-Met) where he stresses that “the location of the soul in the body [. . .] cannot be determined [his emphasis] [. . .] I cannot feel the place in the body where the soul resides.” (AA 28, 281) Yet, despite this epistemological restriction, Kant puts forward an argument that sounds like he is alluding to the contemporary idea of supervenience, which posits the ground of all sensations in the brain. It is worth quoting this passage in full (Kant’s emphases):

> But the cause of all sensations is the nervous system. Without nerves we cannot sense anything outer. But the root of all nerves is the brain; the brain is accordingly aroused with each sensation because all nerves concentrate themselves in the brain; accordingly, all sensations concentrate themselves in the brain. Thus the soul must put the *seat of its sensations* in the brain, as the *location of all conditions* of the sensations. But that is not the location of the soul itself, but rather the location from which all nerves, consequently all sensations as well, arise. [. . .] When, e.g., I hold a finger to the fire, then I experience pain in it; but in the end all sensations from every particular part of the body are concentrated in the brain, the stem of all nerves; for if the nerves from one part of the body are cut, then of course we feel nothing from that part. Accordingly, the principle of all sensations must be in the brain. [. . .] When we imagine a position in the brain which is the first principle of the stem of the nerves where all nerves run together and end in one point, which is called the seat of the senses <sensorium commune>, but which no physician <medicus> has seen, then the question arises, does the soul reside in this seat of the senses <sensorio commun>? Has it taken up a little spot there from which it directs the
whole body, somewhat like an organist can direct the whole organ from one location; or does it have no location at all in the body, so that the body itself is its location? Granted, if the soul took up a little spot in the brain where it plays on our nerves as on an organ, then we could believe that if we had gone through all the parts of the body we ultimately would have to come upon this little spot where the soul resides. Now, if one took away this little spot, the whole human being might still be there, but the location would be lacking where the organist is supposed to play, as though on an organ: but this is thought very materialistically.

(V-Met, AA 28, 281–282)

I want to highlight two impressive features of this passage and of Kant’s engagement with Sömmerring’s proposal in the present context. Firstly, these passages in effect anticipate Dennett’s (1991) objection against what he calls ‘Cartesian materialism’, a position allegedly shared by many contemporary neuroscientists who try to identify certain brain areas or processes as being causally responsible for (or identical with) consciousness. The terminological contrast between a local and a merely virtual presence of the mind in the brain has a very modern ring to it, considering Dennett’s characterization of the mind as a “virtual machine implemented in the parallel architecture of a brain” (Dennett 1991, p. 210), indeed anticipating the functionalist view of the mind. It is difficult to determine, though, whether Kant’s use of “virtual” in his discussion of Sömmerring’s proposal is akin to Dennett’s.

Secondly, Kant even engages with Sömmerring’s specific proposal concerning the liquid contained in the brain’s ventricles, expressing his “great scruple” that this candidate substrate is not organized. Only something having some sort of organization or “purposive disposition of its parts” could serve to locate the mind. This is reminiscent of Kant’s own groundbreaking and very influential discussion of organisms as natural purposes, i.e., self-producing and self-organizing beings in his Critique of the Power of Judgment. In contrast to a mere mechanical organization, Kant considers what he calls a “dynamical organization” to be crucial for the mind. Again, what he means by this is not specified any further in the passage quoted, but it can be illuminated by his discussion of the contrast between mechanistic and teleological explanation in the third Critique. This discussion of the immanent purposiveness of living organisms has inspired generations of philosophers, leading up to the present-day development of so-called “enactive” approaches to the mind (Varela et al. 1991, Weber/Varela 2002, Thompson 2007).

6 For a recent installment of this controversy see the exchange between neuroscientists Alan Hobson and Karl Friston on the one hand and philosophers Krzysztof Dolega and Joe Dewhurst on the other (Hobson/Friston 2014, Dolega/Dewhurst 2015, Hobson/Friston 2016). Schlicht/Dolega (2021) discuss the prospects of the predictive processing framework as a guide of this search for neural correlates.
2.3 Kant and Enactivism

In one of his last texts, Francisco Varela acknowledges his debt to Kant’s ground-breaking discussion of organisms for the development of the ‘autopoietic’ or ‘en-active’ conception of cognition in the early 1990s (Weber/Varela 2002). Together with Brooks’ (1991) work in robotics, the enactive-embodied approach to cognition challenged both the representationalist paradigm and its explicit separation of perception from action in the traditional ‘sandwich conception’ (Hurley 1998) of cognition in favor of a dynamic view. In contrast to a traditional linear progression from sensory input via cognitive computation to action, enactivism conceives of perception and cognition not simply as functional brain states but as entangled and intertwined embodied activities of whole organisms (agents, systems) that can be explained without appeal to mental representations (e.g., Varela et al. 1991, Noë 2004, Chemero 2009, Hutto/Myin 2013, Gallagher 2017). Indeed, in this framework, the equivalence of intentionality and mental representation is no longer taken for granted (Schlicht 2018). Applied to our example used above, perceiving a coffee mug not only requires multiple actions like eye-, head- and body-movements (gaze turning etc.); perceiving is in the service of detecting action possibilities (like grasping) from the start (Gibson 1979).

All enactivists subscribe to what Thompson (2007, p. 128) calls the “deep continuity of life and mind”, i.e., the claim that the organizational features of mind are an enriched version of those of life (Noë 2009, p. 41; Colombetti 2013, p. xvi; Gallagher 2017, p. 102; Di Paolo et al. 2017, p. 3, 178). In an evaluation of Kant’s influence on current cognitive science, this is the crucial aspect of enactivism. At the heart of this conception is the notion of autopoiesis (Maturana/Varela 1980). An autopoietic system – the minimal living organization – is one that continuously produces the components that specify it, while at the same time realizing it (the system) as a concrete unity in space and time, which makes the network of component production possible (Weber/Varela 2002, 115). In his second Critique, Kant conceives of organisms as ‘self-organized’ and ‘self-producing’, i.e., autopoietic, systems that cannot be explained in purely mechanistic terms, but which we have to ‘make intelligible’ to us by relying on teleological principles that are not part of natural science but borrowed from practical contexts. Impressed by certain animals’ (e.g., zebra fish, salamanders) ability to regrow damaged or even severed body-parts (Simon 2012), Kant discusses examples to demonstrate that animals exhibit a certain form or organization that, if conceived merely as the result of blind mechanistic causal processes, appears completely contingent. Yet, “since reason must be able to cognize the necessity in every form of a natural product if it would understand the conditions...
connected with its generation”, our understanding must borrow the concept of final cause to make sense of this organization.

This leads Kant to his conception of organisms as “natural ends”, i.e., as natural products and as ends at the same time. This looks like a contradiction, since the notion of an ‘end’ or ‘purpose’ – being a “stranger” (KU, AA 05: A390) in natural science – must be projected into nature for the sake of an understanding of (some of) its products. Unlike a watch, the parts of an organism, its organs, must be taken to produce themselves rather than being produced by an external power, and they arrange themselves in relation and mutual dependence to each other. Analogously, unlike a watchmaker, in the case of organisms the guiding idea is not to be found outside the product (the watch), but within it (the organism itself). “An organized being is thus not a mere machine, for that has only a motive power, while the organized being possesses in itself a formative power, and indeed one that it communicates to the matter, which does not have it (it organizes the latter)” (KU, AA 05: A374). Thompson (2007, p. 62) refers to this formative power as “circular causality”, i.e., a causal dependence which goes two ways: on the one hand, the features of the whole (organism) are determined by its parts (organs); on the other hand, the local interactions of the parts (organs) are determined by the whole (organism). But as Kant makes explicit in the third Critique, this assumption is to be taken only in an epistemological sense, i.e., we only regard it as if organisms were possible only through reason, since as natural products they must come about through purely mechanistic causes, and thus be amenable to a mechanistic explanation. We cannot prove that organisms indeed exhibit this formative power, since we cannot acquire an intuition of it.7

Francisco Varela regarded Kant’s position as important, because he took Kant to have “developed the possibility of a third way between a strong teleology and a brute materialism” (Weber/Varela 2002, p. 99). Varela acknowledges Kant’s insight but considers his position “unstable” and in need of revision “on the basis of modern developments of biological research and thinking”. According to Weber and Varela, Kant’s conception of an organism as a self-organized and self-producing being is closely analogous to the definition of an organism in Varela’s own theory of “autopoiesis”. In this view, biological autonomy and individuality warrant the assumption of an “intrinsic teleology”, to the effect that “organisms are subjects having purposes according to values encountered in

7 At this point it is important to emphasize that Kant, unlike some of the idealist philosophers after him, does not regard humans to be capable of an intellectual intuition in contrast or in addition to a merely sensory intuition. See Grüne (2009) for more detailed discussion on Kant’s theory of intuition.
the making of their living” (Weber/Varela, p. 102). The theory of autopoiesis as a theory of living systems is supposed to help naturalize Kant’s original theory of organisms. The question of whether Kant’s epistemic and critical position on this issue of teleology or Varela’s naturalistic theory of autopoiesis is warranted, is beyond the scope of this review. But this illustrates how Kant’s philosophy of biology left a footprint with wide-ranging implications in the historical development of cognitive science.

One particular implication of the autopoietic approach to cognition and the mind-life continuity thesis is that all organisms may exhibit at least some basic form of cognition, whereas such views have a problem allowing for genuine cognition in artificial systems. In contrast to more traditional cognitivist approaches, the possibility of cognition in ‘simple’ biological systems has recently been taken seriously with respect to organisms such as bacteria (Ben Jacob et al. 2006), plants (Calvo/Keijzer 2011, Calvo et al. 2020, Mancuso 2018, Sims 2019), and slime molds (Vallverdú et al. 2018), for example. Whether Kant would have regarded the life-mind continuity thesis as credible must be left open here, although Nunez (forthcoming), drawing on the Critique of the Power of Judgement (§65), argues that Kant would have had to at least ascribe desires to plants on the basis of how they move and on how Kant himself treated the notion of being alive.

Developmental biologist Michael Levin somehow takes this story full circle by arguing that we should apply the computational approach not only to animals with brains and nervous systems, but also to simple organisms without a brain. Rather than continuing to contrast the brain with the rest of the body (even in the so-called embodied cognition research program, see Shapiro 2011), Levin invites us to consider the body as performing calculations as well, so as to overcome the traditional life vs. machine dichotomy and according to an updated notion of ‘machine’ (Bongard/Levin 2021). Puzzled by an organism’s formative power, Levin speculates that cells and tissue may exhibit some basic forms of memory and action, using bioelectricity to communicate and decide or plan development (Levin et al. 2021, Pezzulo et al. 2021). For example, he succeeded in ‘reprogramming’ a planarian worm to grow a second head in place of its tail which he had cut off. What he’d done was to change the bioelectrical signals or ‘code’ which would normally have led to the growth of a new tail. Levin’s work suggests a convergence between biology and computer science and is thus highly relevant for the future of artificial intelligence.

8 For further details and discussion see Thompson (2007); cf. Schlicht (2011) for a critical evaluation of Thompson’s account.
2.4 Kant and Predictive Processing

Major developments in machine learning also heavily inspired recently popular predictive processing models of the brain which are taken to provide “the first truly unifying account of perception, cognition and action” (Clark 2016, p. 2) by conceiving of the brain as a prediction machine. This view implicates a delicate balance between bottom-up and top-down processing, in contrast to traditional serial bottom-up processing accounts: Perception and cognition are defined in terms of the brain testing hypotheses about the (sources or causes of) incoming sensory stimulation; hypotheses are generated by a hierarchical generative model of the world and constantly updated in response to prediction error signals (Friston 2010, Hohwy 2013, Clark 2016, Metzinger/Wiese 2017). To return to our example, perceiving the coffee mug is a process already informed by underlying brain processes that constitute a set of more or less likely expectations about sensory input and its causes. These expectations are constantly compared to the actual incoming sensory information, resulting in prediction errors (deviations) that are processed in the brain. The traditional picture of the brain using incoming sensory information to build up a representation of the world is thus turned upside down, since the new picture holds that “the rich representation of worldly states of affairs is signaled in the top-down predictions of sensory input, maintained by the perceptual hierarchy in the brain” (Hohwy 2013, p. 47).

Link Swanson (2016) argues that this most recent paradigm in cognitive science also has roots in Kant’s philosophy and tells a convincing story tracing back this influence via Helmholtz’s thesis of perception as unconscious inference, which in turn was a primary source for Friston’s (2005; 2010) original proposal regarding predictive processing as a unified brain theory. The radical reversal of processing (top-down hypothesis-testing rather than bottom-up model-building) characteristic of predictive processing finds an analogue in Kant’s so-called Copernican revolution with its combination of intuition (providing the material) and concepts (generating an understanding of what’s perceived), presenting “us with a view of perception as a Kantian in spirit, ‘spontaneous’ interpretative activity, and not a process of passively building up percepts from inputs” (Gładziejewski 2016, p. 574). But Swanson also links more specific concepts from the predictive processing story to specific analogues in Kant’s theory – e.g., generative models and schemata, which are both heavily informed by intuitions as well as concepts in the process of object recognition. Indeed, it’s striking that both Clark and Hohwy choose a starting point that sounds very familiar to Kantians, only formulated from the perspective of the brain whose task, “when viewed from a certain distance, can seem impossible: it must discover information about the likely causes of impinging signals without any form of direct access to their source”
(Clark 2013, p. 183). Put this way, the central issue is understanding causation, i.e., understanding relations between worldly causes and sensory inputs. Assuming a Humean framework, this is impossible according to Kant, who posits an innate conceptual machinery (the categories) that must be applied to sensory input in order to enable such understanding.

But only Hohwy’s interpretation of the predictive processing framework is internalist like Kant’s. Indeed, Hohwy argues that the prediction error minimization theory “reveals the mind to be inferentially secluded from the world”, showing that strong embodied views of cognition and mind should be rejected in favor of “a more old-fashioned, skepticism-prone view of the mind-world-relation” (Hohwy 2014, pp. 259–260). Thus, Beni (2018) complains that Swanson’s reconstruction only holds true for Hohwy’s version of predictive processing, while it ignores the dominant embodied and action-oriented versions put forward by Clark (2016) and Bruineberg and Rietveld (2014) or Bruineberg, Kiverstein and Rietveld (2016), which are much more inspired by Gibson’s (1979) ecological psychology than by Kant’s transcendental idealism. Whether the embodied variety of predictive processing is tenable and coherent, given the epistemic starting point it shares with Hohwy’s version, cannot be pursued further in this paper. In any case, the view that predictive processing is rooted in Kant’s view of the mind must thus be taken with a grain of salt, just like the idea that assimilates Kant’s view of the mind to functionalism.9

With respect to the alleged roots of functionalism, enactivism, and predictive processing in Kant’s philosophy, it is important to keep in mind that these different paradigmatic backgrounds take different stances towards the relation between cognition, intentionality, and representation, and propose different explanatory strategies in cognitive science. It seems unlikely that Kant would have subscribed to all of these views at once, given the opposition (and genuine incompatibility) of some of the contemporary stances.10 While classical cognitivism alludes to mental representations, embodied enactivism eschews them (Gallagher 2017, p. 7). While the former is based on functionalism and explicitly allows for the possibility of cognition in artificial systems, the latter is based on a strong continuity of life and mind, making this possibility problematic.

---

9 Anderson & Chemero (2019) indeed criticize Clark for trying to embrace both an embodied-enactive view of the mind and the predictive processing architecture, simply because it seems to be somewhat rooted in Kant’s framework, which they take to have been overcome by an embodied perspective on the mind.

10 See Di Paolo, Thompson, and Beer (2021) for a statement on the incompatibilities between the free-energy approach (underlying the predictive processing framework developed by Hohwy and Clark) and enactivism.
2.5 Interlude: Aspects of Kant’s Account of Cognition

As we saw, at the heart of the predictive processing approach is the project of understanding causal relations. This is also the recurrent theme linking Hume’s empiricism with Kant’s transcendental idealism and his ‘Copernican Revolution’. Hume recognized that an understanding of a causal relation between A and B cannot be grounded in sensory input alone, since this does not provide us with a connection between events A and B but only with their temporal succession, yielding his skepticism about an understanding of causation. Kant followed Hume in his assessment of the inadequacy of sensory experience in accounting for an understanding of causal relations; at the same time, he was willing to borrow a priori concepts from the rationalists and claim that it is the faculty of understanding itself which is the source of a system of concepts that provide the necessary unification. It is now worth reminding the reader of the core of Kant’s theory of cognition, prior to the discussion presented in the second part of the paper.

In two very instructive papers, Marcus Willaschek and Eric Watkins (Watkins/Willaschek 2017, Willaschek/Watkins 2020) outline the complex usage of the notion of cognition in Kant’s works. The most prominent usage is what they call cognition in the narrow sense, which requires a unification of intuition and concept, i.e., the combination of sensory receptivity and spontaneity of the understanding. Cognition in the broad sense, by contrast, allows for several “degrees of cognition”, sketched in different, yet not necessarily incompatible, ways in the so-called Jäsche Logik (AA 16: 64–65) and in the Critique of Pure Reason (KrV: A320/B376). In the latter, they are presented as more or less demanding cases of “representing something”, be it unconsciously, consciously, through perception, understanding or reason, with or without concepts or intuition being involved. The most basic degree of cognition is “to represent something”, without any further conditions; the highest or most complex degree is to “comprehend something” through reason and a priori. Importantly, in Kant’s taxonomy, cognition does not imply truth or assent, and is therefore to be distinguished from the notion of knowledge (Willaschek/Watkins 2020). Taken in the broad sense, any conscious representation that represents an object counts as a case of cognition, even if the object does not exist (or if it cannot be given in experience).

But despite the variety of dimensions of the concept of cognition, as used by Kant, cognition in the narrow sense is singled out as “cognition in the proper sense” (KrV: A78/B103) and this is the notion that will concern us here. Cognition in this sense can be described as a “conscious representation of a given object
and of (at least some of) its general features” (Watkins/Willaschek 2017, p. 86).
For cognition of an object to obtain, this must be given and a concept must be
applied to it. The former is the task of sensibility, the latter is performed by the
understanding. And such cognition is actively achieved rather than simply hap-
pening by chance, since it is a product – “the mere effect” – of the synthesis per-
formed by the imagination, “without which we would have no cognition at all”,
at least not in this crucial narrow sense. This passage places great emphasis on
the function of synthesis, which is conceived as “the action of putting different
representations together with each other and comprehending their manifoldness
in one cognition” (Watkins/Willaschek 2017, p. 86). An act of synthesis, as such
the beginning of an answer to the problem posed by Hume, “collects the ele-
ments for cognitions and unifies them into a certain content”. Without such a
unificatory process of concept application to a given object, intuitions remain
“blind” and thoughts “empty” (KrV: A51/B75-76). Kant therefore stresses that
if we are interested in “the first origin of our cognition”, we have to focus on
synthesis.

What the conception of degrees of cognition in the Jäsche Logik and the
“progression” passage in the Critique of Pure Reason have in common is the
idea that cognition in the narrow sense presupposes consciousness. In the Jä-
sche Logik, where Kant outlines a gradual concept of cognition, this idea is
found in the fourth degree, defined as “to be acquainted with something with
consciousness, i.e., to cognize it” (AA 16: 65), whereas in the “progression” Kant
develops it as follows: “The genus is representation in general (repraesentatio).
Under it stands the representation with consciousness (perceptio). A perception
that refers to the subject as a modification of its state is a sensation (sensatio);
an objective perception is a cognition (cognitio)” (KrV: A320/B376). Taking the
progression seriously, Tolley (2020) argues that Kant classifies sensing, intuit-
ing, perceiving and mere thinking as “lying earlier” than, and providing condi-
tions for, cognition, while still considering cognition as being placed on a
“psychologically elementary level” compared to knowledge, understanding,
and explaining. In contrast to Watkins and Willaschek, Tolley argues that
Kant’s concept of cognition is unified rather than equivocal.

Without intending to settle this dispute with respect to Kant’s use of “cog-
nition”, all sides agree that he emphasized cognition in the narrow sense,
where the other candidates fall under the umbrella of the concept of cognition
as it is used in contemporary cognitive science. And since cognition in the narrow
sense, in Kant’s view, is a “distinctive form of consciousness of a real object by
way of a specific kind of combination of representations” (Tolley 2020, p. 3217),
consciousness is a condition of cognition in this proper sense (I will return to this
point in the last section of this paper). By Kant’s lights then, for an artificial system
to be capable of cognition in the narrow sense, it would have to be capable of conscious-ness as well. This is certainly not a view of cognition that is widespread among contemporary cognitive scientists. Proponents of the predictive processing approach to perception and cognition also do not hold that these processes require consciousness, although they often claim that the framework can also be applied to explain consciousness (Hohwy/Seth 2020, Clark 2019). But even if an artificial system may not be conscious in the relevant sense, it may still be capable of cognition in the broad sense. That is, it may be said – minimally – to have representations of something or other.

With these reminders of Kant’s view of cognition, we can now turn, in the second part of this paper, to more recent developments in artificial intelligence, namely, the ascent and success story of deep learning architectures that led to the recent AI spring. As I will try to show, this fascinating development raises interesting philosophical issues about the nature of perception, learning and understanding and about the more general question of empiricist vs. rationalist approaches.

3 The Potential and Limitations of Deep Learning

After a series of dark winters, AI research has made considerable progress, pushed forward by the advent of so-called “deep learning architectures” (LeCun/Bengio/Hinton 2015; Buckner 2018/2019, Sejnowski 2018). This has been – so far – the result of a development in which the connectionist approach to AI superseded the “Good Old-fashioned AI” (or GOFAI-) approach. Current discussions of AI often focus on software that does not only process fixed programmed algorithms, but can be trained using algorithms, such that part of the process can be developed by the software itself. This machine-learning approach to AI is one among many and includes supervised, unsupervised and reinforcement learning. Deep Learning on the basis of artificial neural networks is currently the most promising and most widely discussed (and used) approach, which is why we will focus on it here.

The crucial difference compared to traditional GOFAI-approaches is that these neural networks are inspired by the organization of the human brain – more specifically, the layered architecture of the visual cortex – while the main difference compared to their historical precursors, the connectionist networks from the 1980s and 1990s, is the number of layers of simulated neurons. Whereas classical networks only consisted of an input layer, one hidden layer and an output layer, deep neural networks are deep in the sense that there are many more
than one hidden layer, indeed there are numbers reaching hundreds of layers. This increases their computational power exponentially, enables them to represent even abstract features of the environment and is taken to be largely responsible for their recent success in many applications. Thus, although this new phase already started in the 1980s, researchers only developed computers with the necessary computational power in the late 2000s. The nodes of the network are connected – just as real neurons are connected via dendrites – and the connections between them have different weights. The larger the weight between A and B, the greater the influence of A on B, and vice versa (since weights are symmetric).

Using cats, neuroscientists and Nobel laureates David Hubel and Torsten Wiesel (1962) discovered that light of different wavelengths activates cells in the back of the eye and that this activation is then processed via the optic nerves into the brain, ending up in the hierarchically organized series of layers of neurons in the visual cortex. Neurons in different layers have specific preferences (or receptive fields) and thus detect increasingly complex features, from edges via simple and complex shapes to whole objects, like faces.\(^\text{11}\) The *nodes* of the network are like simplified, formal neurons. The input layer provides the data for the network – images, spoken words, hand-written digits, games –, whereas the output layer produces the desired results, e.g., a classification of an image or object, a number or word. In between, multiple hidden layers perform calculations that produce this result:

An image, for example, comes in the form of an array of pixel values, and the learned features in the first layer of representation typically represent the presence or absence of edges at particular orientations and locations in the image. The second layer typically detects motifs by spotting particular arrangements of edges, regardless of small variations in the edge positions. The third layer may assemble motifs into larger combinations that correspond to parts of familiar objects, and subsequent layers would detect objects as combinations of these parts. The key aspect of deep learning is that these layers of features are not designed by human engineers: they are learned from data using a general-purpose learning procedure. (LeCun/Bengio/Hinton 2015, p. 436)

Several major factors are important for their performance:

1. First, as the name suggests, these networks are able to *learn* and can therefore be *trained* on the input data; they are not pre-programmed (although the programmer chooses the input data). Learning takes place by adjusting the weights

\(^{11}\) Hubel and Wiesel dubbed them “simple” and “complex” cell types respectively. This inspired Fukushima (1980) to develop a new kind of network, the “Neocognitron” which was supposed to demonstrate this stronger computational power.
according to sensory feedback. These networks start with arbitrary weights and adjust them in the course of a training phase in which they are bombarded with data. If the goal is to learn recognizing objects, the inputs will be images; if the goal is to learn playing games, the input will be games of this sort. And so on.

(2) This is the second important factor: Using internet databases such as image-net, the training set for a given network can consist of millions of examples, e.g., millions of images of dogs and cats, or millions of games of Go – many more dogs and cats and games than any human being could encounter or play in their lifetime. Note well: the point is not that since DNNs can rely on so much data, they have a significant computational advantage when compared with humans; the point is that they must rely on so much data to achieve this significant level of performance. Children, by contrast, can learn very quickly from just a few examples (Carey 2009). That’s an important difference. Since the real world does not come as neatly labeled as suggested by a supervised learning training set for networks, this cannot be the route to mimic human learning or understanding. It is different from the very start. But that does not preclude us from considering the procedure “intelligent” or as an instance of “cognition”, since these phenomena may allow for multiple realizations.12

Yet, if the images come already labeled (this is a dog, this is a cat) – which is the most common method of machine learning – the network will eventually learn to produce confident results (outputs) in recognizing dogs and cats. In general, the output does not consist of a single answer, but comes as a “vector of scores, one for each category” where the goal is to get the machine to give the desired category the highest score (LeCun et al. 2015, p. 436). For example, if the input is an image of a dog, the network might spit out 70% dog, 20% fox, 10% cat, i.e., outputs with different confidence ratings. Eventually, after initially making many errors, performance increases because these errors are processed using so-called ‘backpropagation’: it calculates the difference between the intended and the actual output and sends this error signal (this is not a dog) back through the hidden layers of the network. By adjusting the weights along the way, the network can perform better the next time it encounters this image.

Although it isn’t clear whether there is a biological analogue to this process of backpropagation, it works well for these networks. Since what a network can ‘recognize’ on each layer is not pre-programmed, it must find out about the most salient and characteristic features that are central for the task (of recognizing dogs, say). This is important for the test phase in which the network is

12 See Buckner (2020) for rebuttals to similar criticisms against deep learning networks.
supposed to classify and recognize with high confidence new objects (more cats and dogs) which weren’t in the training set. So far, DNNs do not make it intelligible to us how they reach a decision when they recognize an object with high confidence, for example, 60% dog, 30% cat, 10% fox. The most prominent artificial neural networks today – convolutional neural networks, or ConvNets (Mitchell 2020, pp. 73–88) are named after the operation leading the DNN to yield a certain output: convolution.\(^\text{13}\) Convolution is a mathematical procedure that works like a filter that slides across an image and creates a layer of features across this image (Sejnowski 2018, p. 130–131). It thereby determines whether a certain portion of an image, say, a set of pixels in a grid, contains or signifies a certain feature and then assigns a certain numeral to that part of the grid. Repeating this procedure for several layers covers ever increasing portions of the image and detects ever more abstract features, thus corresponding to Hubel and Wiesel’s ‘simple cells.’

The network is still relying on human expertise in the form of feedback (labels) about its results. That makes the learning process “supervised”. The inputs are fixed and the results are determined; the network must learn to get from A to B using only its own resources, simulated neurons in multiple layers connected by different weights. After having received the input image, the network performs its layer-by-layer calculations and finally produces an output. This can be formulated as a certain degree of confidence (between 0 and 100%) regarding every image and category. Some networks have already achieved more than 90% accuracy in the image-net competition (Mitchell 2020, p. 101). Nevertheless, their performance is limited, since they can only succeed or fail in categorizing an input-picture (or word) correctly but they cannot produce any new insights. This seems possible in unsupervised learning when the result (the label “dog”, say) is not given but found by the network by associating and clustering certain patterns with each other. For example, the network might be able to detect words in social media feeds which are used more frequently than others or might recognize that customers who bought product A also often bought product B, which can then be recommended to new customers who bought A.

(3) A third and peculiar aspect of DNNs concerns the kinds of errors they make and how different these errors are from the kinds of errors humans make.

---

\(^\text{13}\) According to LeCun, Bengio and Hinton (2015, p. 439), ConvNets (or DCNNs) proved to be easier to train and “generalized much better” than other networks, cf. LeCun et al. (1990). ‘Pooling’ is a procedure that is seen as corresponding to convolution and similar to Hubel and Wiesel’s ‘complex cells’, since it aggregates each feature over a region and guarantees for invariance.
Since humans also make mistakes – being subject to visual illusions, for example –, such networks need not be perfect in their performance. But it is instructive how easily they can be fooled:

While they also get confused by images containing multiple objects, unlike humans they tend to miss objects that are small in the image, objects that have been distorted by color or contrast filters the photographer applied to the image, and “abstract representations” of objects, such as a painting or statue of a dog, or a stuffed toy dog.

(Mitchell 2020, p. 105)

Moreover, and most disconcertingly, DNNs are easily duped and fooled, both by intentional manipulation of the input data and by new situations in the real world that the network is not sufficiently prepared for. Thus, a self-driving car’s autopilot mode got confused by salt lines which had been laid out on a road in anticipation of a storm, since they looked just like lane markings – an unlikely yet possible situation. Mitchell (2020, p. 120) reports results from her colleague Will Landecker who had trained a network in classifying images into “contains an animal” and “does not contain an animal”. But the test phase revealed that the network had classified all photos with a blurry background as containing an animal since there was a high correlation between a macroscopic picture taken of an animal and the photo having an otherwise blurry background. That is, the network ‘overfitted’ to its training set and thus failed to accurately predict future data that are slightly dissimilar but nevertheless belong to the same relevant category. Another, rather embarrassing and inexcusable because discriminatory, mistake happened to the Google Photos App when it labeled a selfie taken by an African-American couple as “Gorillas” (Vincent 2018).

DNNs can be fooled more systematically using so-called “adversarial examples”. These are images which have been intentionally distorted by making very small changes that the human eye cannot detect but lead a DNN to classify the object depicted on it in an arbitrary manner, even though it had correctly classified the original image before. For example, a lion was now classified as a library, both times with high confidence (Szegedy et al. 2015). Subsequently, Nguyen, Yosinski and Clune (2015) showed that it is possible to produce copies of images showing an object A where the copy contains differences which are unrecognizable to the human eye and yet allegedly recognized as showing another object B with 99% confidence by a DNN. This seems to show that not only do DNNs learn very differently than humans – “at the most specific grain of detail, DCNNs [Deep Convolutional Neural Networks, T.S.] and human perceptual cortex do not produce exactly the same phenomena” (Buckner 2018, p. 28)\(^\text{14}\) – they also cannot be

\(^{14}\) Generative Adversarial Networks (GANs) can be seen as a reaction to this problem. Here, one network’s task is to recognize and classify images correctly, while another one generates
trusted. Thus, it remains obscure why ConvNets work as well as they do.\textsuperscript{15} This opacity of the learning and decision-making process makes it difficult to understand what and how such networks learn. Mitchell (2020, p. 132) concludes that “something very different from human perception is going on”.

### 3.1 Philosophical Interpretations of Deep Neural Networks

As we saw in the first part, one crucial aspect of Kant’s theory of cognition is that he posits a balanced interaction between bottom-up and top-down processing, in his terminology between intuition and concept. A second, yet different, aspect of his approach is the positing of a priori contributions to cognition, i.e., contributions that are independent and systematically prior to experience or learning. With respect to our understanding of causal relations, for example, Kant shared the same starting point with Hume, stressing that it cannot be conceived as a “direct consequence of data-driven learning” (Butterfill 2020, p. 93). However, contrary to Hume’s skepticism, Kant concluded that there must be a contribution to understanding that is not learned which is often identified with being “innate”.\textsuperscript{16}

Considering the preceding paragraphs, DNNs also raise a number of interesting issues concerning perception, classification, abstraction, conceptual learning, and understanding, and also concerning the debate about innate vs. learned against the background of the controversy between empiricists and rationalists. Buckner argues that in today’s debates, the question is no longer whether the mind starts out as an unstructured ‘blank slate’ but whether categorical representations “are due mostly to domain-specific or domain-general cognitive mechanisms” (Buckner 2018, p. 3). A typical example for a domain-specific cognitive mechanism is Chomsky’s universal grammar, which constitutes a language acquisition device underlying our learning of all natural languages; it is domain-specific since it only pertains to language. By contrast, one may posit only one domain-general all-purpose learning device allowing one to acquire knowledge across domains (as done, for example, in Skinner 1957, the book criticized by Chomsky 1959). Crucially, opponents in the debate would not count the
data like those in the training set. In this competitive way, the two networks can improve on their performance.

\textsuperscript{15} This feature of DNNs gives rise to the research program of “explainable AI” which aims at making such networks intelligible (see also Sejnowski 2020).

\textsuperscript{16} Samuels (2004, p. 139) criticizes this characterization for being uninformative.
latter as evidence for a nativist position, since, as Long (ms., p. 3) argues, everyone agrees that learning requires that *something* be innate. He proposes, following Margolis and Laurence (2013), to frame the controversy in terms of this contrast, with nativism holding that cognition (in a given domain) requires domain-specific mechanisms, and empiricism holding that (for any domain) domain-general mechanisms are sufficient. The questions we are facing then are the following: Given that DNNs do not start from scratch, do they require domain-specific mechanisms or can they make do with domain-general ones in order to achieve general intelligence? (2) Does Kant’s system of categories constitute a domain-general or a domain-specific cognitive mechanism?

Regarding question (1), Long (ms.) has usefully framed the development of artificial intelligence in terms of this opposition and formulated more fine-grained theoretical options. To keep things as simple as possible, we will focus on only two of them: “Necessity Nativism is the claim that necessarily, a human-level AI system will be a nativist system” (Long, ms., p. 6). That is, in this view general intelligence requires nativist (domain-specific) mechanisms. By contrast, *Possibility Empiricism* “is the claim that it is possible for a human-level AI system to be an empiricist system” (Long ms., p. 6); i.e., general intelligence, in this view, does not require domain-specific mechanisms but can be acquired by relying on domain-general mechanisms alone. Long argues that “empiricist human-level AI is at the very least possible” (Long, ms., p. 1). That is, Long belongs to a group which we may dub “optimists”. Optimists claim that developers may overcome the obstacles that current AI systems face compared to human-level understanding without having to rely on domain-specific mechanisms. Pessimists, by contrast, claim that developers will not succeed in building AI systems that can achieve human-level understanding solely by relying on domain-general mechanisms. Domain-specific, i.e., innate mechanisms are necessary for this feat.

(a) Optimists

In a seminal article, LeCun, Bengio and Hinton (2015, p. 436) claim that DNNs are able to “learn representations of data with multiple levels of abstraction”. If that were so, this would be fantastic, since – according to Mitchell (2020, p. 319) – “abstraction, in some form, underlies all of our concepts, even from earliest infancy” and it would therefore open the door for the possibility that DCNNs may acquire concepts and understanding simply from being exposed to data. That is, LeCun, Bengio and Hinton are optimists. So is Buckner (2018), who is impressed by the success of ConvNets and discusses them in the context of an empiricist
philosophy of mind, claiming that they “model a distinctive kind of abstraction from experience”, and thereby “one crucially important component of intelligence – a form of categorial abstraction”, among other components necessary for general intelligence (Buckner 2018, p. 3). He also highlights several core features of DNNs – multiple layers, convolutional filters, and pooling – and argues that “they jointly implement a form of hierarchical abstraction that reduces the complexity of a problem’s feature space [...] by iteratively transforming it into a simplified representational format that preserves and accentuates task-relevant features while controlling for nuisance variation”, i.e., variations that are irrelevant for categorization (size, location etc.). He calls this process “transformational abstraction” (Buckner 2018, p. 18). In his rich and densely argued paper, Buckner nicely presents both Locke’s as well as Berkeley’s and Hume’s somewhat mysterious and unsatisfactory accounts of abstraction, culminating in the puzzle of how the mind can get from specific exemplars to abstract categories (Locke) or from abstract categories to exemplars (Hume). Where does the knowledge come from which details should be left out (Locke) or added (Berkeley, Hume) along the way? While at one point he acknowledges that what he is developing “begins to look more like the theory of abstraction provided by Kant (and contemporary Kantians like Barsalou [...]) who emphasized the need for rules of synthesis to generate a range of specific possible exemplars corresponding to an abstract category” (Buckner 2018, p. 12), he nevertheless argues, along the lines of possibility empiricism, that this challenge may be met by an empiricist account. He is content to have shown that DNNs perform abstractions that vindicate “elements of the Lockean, Berkeleyan and Kantian views”, without committing himself neither to any one of these historical interpretations, nor to the crucial differences between these accounts.

(b) Pessimists

At the time being, it is fair to say that the group of pessimists pointing out several limitations of Deep Neural Networks is larger, or at least louder than the group of optimists (depending on whom you talk to). As Buckner (2018) notes, contemporary rationalists are skeptical about domain-general mechanisms being sufficient. Indeed, Mitchell (2020, p. 132) argues that the main problem of deep neural networks is “one of understanding”. The networks lack the rich background knowledge – about functions of objects (affordances), memories, and context dependent cognition – which informs human perception. She suggests that “humans are endowed with an essential body of core knowledge” (Mitchell 2020, p. 309), appealing to the influential work by Spelke and Carey (1996) which
posits domain-specific core knowledge systems enabling the recognition of objects, agents, numbers, and so on – i.e., concepts like cause, number, object, and agent. The list of features of these systems typically contains “innateness”.

An even more dismissive assessment of what deep neural networks can achieve is that given by Marcus and Davis (2019, p. 145). They agree with Mitchell but go further, objecting that what DNNs provide is just more of the same that was already possible with their precursors. They complain that “machine learning people, for the most part, emphasize learning, but fail to consider innate knowledge” (Marcus/Davis 2019, p. 144). Also appealing to Spelke’s work, they submit that

humans are likely born understanding that the world consists of enduring objects that travel on connected paths in space and time, with a sense of geometry and quantity, and the underpinnings of an intuitive psychology. Or, as Kant argued [. . .], an innate ‘spatio-temporal manifold’ is indispensable if one is to properly conceive of the world.

(Marcus/Davis 2019, p. 145)

Leaving Kant and the question of whether Marcus and Davis’ charge against machine-learning researchers is justified aside for the moment, it should be emphasized that innateness is not a necessary feature of the core systems identified by Spelke and Carey. As Butterfill (2020, pp. 93–103) shows, the evidence for such systems being innate is far from clear, and “poverty of stimulus arguments” have been provided only in the case of syntax (Chomsky 1959), whereas other works in developmental psychology suggest an agnostic position on innateness of core systems. Thus, one may accept the evidence mentioned in favor of a distinction between a limited number of core knowledge systems but nevertheless reject the claim that they can be cited in favor of necessity nativism.

In a similar vein, and with a focus on the goal of developing an artificial system exhibiting general intelligence, computer scientist and philosopher Judea Pearl (2018, p. 10) considers “machines’ lack of understanding of causal relations” as being “perhaps the biggest roadblock to giving them human-level intelligence [. . .] I believe that strong AI is an achievable goal and one not to be feared precisely because causality is part of the solution.”17 “A causal reasoning module will give machines the ability to reflect on their mistakes, to pinpoint weaknesses in their software, to function as moral entities, and to converse naturally with humans about their own choices and intentions.” Yet, despite his optimism, he considers present-day learning machines still only as “sharing the wisdom of an owl”. Despite regular news about rapid advances in machine-learning systems – self-driving cars, speech and face

17 For Pearl, strong AI is the claim that an AI system exhibits general human-level intelligence.
recognition systems, and the like—even recent deep learning networks have only "given us machines with truly impressive abilities but no intelligence. The difference is profound and lies in the absence of a model of reality" (Pearl 2018, p. 30).

In order to illustrate what’s missing compared to the human level of understanding, Pearl sketches a threefold “ladder of causation” (Pearl 2018, pp. 23–52), specifying three levels of cognitive ability that a learner must achieve for a true understanding of causal relations. The first and most basic level, which we share with many animals, consists in detecting regularities through observation. An owl may observe a rat and figure out where it will be next, for example. Such reasoning proceeds merely by association and seeing such regularities enables a cognitive agent to make predictions guided by the question ‘What if I see [. . .]’? Only some of such observations may actually discover causal relations, and the data themselves do not disclose cause and effect. The second cognitive level is characterized by action which enables a cognitive agent to bring about changes in the world. Actions are interventions into the physical causal order. To use Pearl’s example, “seeing smoke tells us a totally different story about the likelihood of fire than making smoke” (Pearl 2019, p. 31). Humans use such interventions all the time, e.g., when taking an aspirin to cure a headache. Tool use in the animal kingdom is an illustration of the range of creatures capable of this cognitive level. The guiding question on this level is: ‘What if I do [. . .]?’ The final cognitive ability, enabling a human-level understanding of causal relations, is counterfactual reasoning. Once the headache is gone, we can ask why and consider the probabilities of different causes, asking, in effect, ‘What if I had done [. . .]?’ An instance of this question is, e.g., ‘What would have happened if I had not taken the aspirin?’ Such thinking opens up new possibilities, taking us beyond data into an imaginary world where some facts, obtaining in the real world, do not hold, or are even contradicted. This hallmark of human intelligence enables the development of scientific theories, art, and improving on our past actions.

The point of all this is that, in Pearl’s view, present-day AI has not yet progressed beyond level one because even DNNs operate entirely in association mode, being fully driven by a stream of data. Recall that being a direct consequence of data-driven learning was how we identified the empiricist position. According to Pearl then, this empiricism is limited. Even the successful computer program Alpha Go only churns through accumulated data, its database consisting of millions of Go games, “so that it can figure out which moves are associated with a higher percentage of wins” (Pearl 2018, p. 29). But this is all that it is capable of, obviously exceeding human memory by far, but not achieving any understanding. By contrast, humans make use of a mental “model of
reality” (Pearl 2018, p. 30, see also Mitchell 2020, ch. 14) which Pearl considers as a necessary ingredient to achieve our level of understanding. While many researchers in AI attempt to solely rely on data for all cognitive tasks, Pearl emphasizes “how profoundly dumb data are about causes and effects” (Pearl 2019, p. 16).

Whether optimists or pessimists may turn out to be correct is an empirical question, and not one to be settled in this paper. At least, the limitations and challenges are more or less known. Cremer (2021) presents a survey of expert interviews on the potential and limitations of deep learning wherein such experts list forty limitations. Success in this area depends on whether the question is if AI systems are supposed to exhibit cognition and intelligence that is like cognition and intelligence in humans or if we would be content with such systems being successful in a sufficiently high number of tasks or in a sufficiently high number of domains, regardless of whether the way they achieved this resembles the way humans do. I highlighted the potential connections to (and relevance of) Kant’s theory of cognition in this context as well as to its borderline position between empiricism and nativism. This was the second question we posed above. On the one hand, given that his system of categories is characterized as being a priori (and thus systematically prior to any experience), his position would be classified as nativist. On the other hand, given that his system of categories can be considered domain-general, his position would be classified as empiricist. Thus, given the different terminological systems used by Kant and contemporary cognitive scientists, it is challenging to formulate a clear statement on this issue that is both true to Kant’s writings and to the way dichotomies are characterized in today’s debates. As a final remark in this review, it is worth introducing a very recent approach to machine learning that specifically alludes to Kant’s view of cognition in its formulation of a cognitive architecture, namely the position developed by Richard Evans (2022, this volume 39–103).

### 3.2 Start Making Sense!

Like Pearl and Mitchell, Richard Evans and his colleagues – computer scientists at Deep Mind, one of the leading companies developing state-of-the-art AI – allude to mental models and add specifications about its ingredients and constraints (Evans et al. 2021). Concerned with the problem of how to make sense of a sensory sequence, they allude to Kant’s theory, since “Kant defines exactly what it means to make sense of a sequence: to reinterpret that sequence as a representation of an external world composed of objects, persisting over time,
with attributes that change over time, according to general laws” (Evans 2022, this volume, p. 40). This “involves constructing a symbolic causal theory that both explains the sensory sequence and also satisfies a set of unity conditions” (Evans et al. 2021, p. 1). More specifically, they postulate the “requirement that our theory exhibits a particular form of unity: the constituents of our theory – objects, properties, and atoms – must be integrated into a coherent whole [. . .] This extra unity condition is necessary, we argue, for the theory to achieve good accuracy at prediction, retroduction, and imputation” (Evans et al. 2021, p. 2). To meet these unity requirements on sense-making, Evans suggests interpreting Kant’s first Critique as providing a cognitive architecture, specifically “as a precise computationally-implementable description of what is involved in making sense of the sensory stream” (Evans 2022, this volume, p. 40). That is, according to Evans, it is possible to capture Kant’s cognitive architecture in rigorous algorithmic form and implement it in a machine in order to test it in experiments.18 Although not all details of Kant’s account can thereby be captured, the gain is a detailed and precise description on the level of a computer algorithm. That is, Kant’s a priori psychology here forms the template for a machine-learning system which requires translating the various faculties that are involved in cognition in the narrow sense and their interaction into one program. In terms of concrete results, these are the understanding – with its capacity to form judgements – corresponds to an unsupervised learning program, the power of judgement, which subsumes intuitions under concepts, is implemented as a binary neural network, and the imagination which is responsible (and indispensable) for the connections between intuitions (productive synthesis) in terms of a set of non-deterministic choice rules (Evans 2022, this volume, p. 95). The fourth and final condition is sensory intuition, which provides the input for the cognitive architecture.

In a more recent article, Evans et al. (2021) describe a particular computer system they call “Apperception Engine”, designed to perform an “unsupervised program synthesis” (Evans et al. 2021., p. 2) and to implement the various

---

18 Moreover, he takes this to yield original rather than merely derived intentionality. The latter presupposes another system with original intentionality (a human being) that can confer its intentionality on the system in question (by interpreting its symbols, for example). Notoriously, Searle (1980) postulates that only a biological system is capable of original intentionality, while all artificial systems will always remain capable only of derived intentionality. Evans (2017, p. 43) argues, by contrast, “that a computational agent built to satisfy a Kant-inspired cognitive architecture is capable of achieving original intentionality. It doesn’t matter what it is made of as long as it achieves the necessary structural organization.”
faculties in one unified system. Delving into the rich details of this implementation is beyond the scope of this review and must be left for another occasion, but the readers may consult Evans’ contribution to this volume themselves (Evans 2022, this volume, ch. 2). This requirement explicitly exceeds the typical empiricist approaches that are purely data-driven, as criticized by the “pessimists” such as Pearl, Mitchell and Marcus & Davis (see above). But it does not mean that Evans thereby takes his Aperception Engine to constitute a nativist system, as demanded by Marcus and Davis. With respect to the debate between optimists and pessimists, Evans objects to Marcus’ interpretation of Kant as a nativist, because it is important what is taken to be innate. That is, it makes a difference whether one claims that concepts are innate or faculties (capacities) whose application produces such concepts. Kant allegedly did not conceive of the categories as innate concepts: “The pure unary concepts are not ‘baked in’ as primitive unary predicates in the language of thought. The only things that are baked in are the fundamental capacities (sensibility, imagination, power of judgement, and the capacity to judge) [. . .]. The categories themselves are acquired – derived from the pure relations in concreto when making sense of a particular sensory sequence” (Evans 2022, this volume, p. 74). Evans follows Longuenesse (2001), who grounds her interpretation in a letter Kant wrote to his contemporary Eberhard; in it, he distinguishes an “empirical acquisition” from an “original acquisition”, the latter applying to the forms of intuition and to the categories. Evans is right in saying that, as far as the cognition of an object is concerned – like the “I think” – the categories come into play only by being actively (spontaneously) applied through the understanding, and can thus be derived, if you will, through a process of reverse engineering which reveals that they have to be presupposed in the first place, being a transcendental condition of experience. But this is compatible with the claim that, given their a priori status (and given that they can be applied also in the absence of sensory input, though not to yield cognition in the narrow sense but still cognition in the broad sense, as characterized above), “they have their ground in an a priori (intellectual, spontaneous) capacity of the mind” (Longuenesse 2001, p. 253). In contrast to Evans, Barsalou (1999, p. 581) firmly categorizes Kant as a nativist, arguing that Kant “assumed that native mechanisms interpret and organize images”. If such mechanisms are supposed to be the categories, then this interpretation speaks against Kant’s elaboration in his letter to Eberhard, but we need not settle this issue here.

---

19 They formulated the unity conditions and chose the name for the computer program in honor of Kant’s background theory in the *Critique of pure reason* (Evans et al. 2021, p. 8, n. 17).
One way the categories are applied is described in the schematism chapter of the first *Critique*. Thus, when the question is how we “form the general idea of a triangle when we have only been exposed to a series of particular and idiosyncratic exemplars” which we are forced to “unify” (Buckner 2018, 8), then Kant’s answer can be framed in terms of the combination of sensory and conceptual representations, arguing that unificatory concepts (categories) are “schematized” when combined with a series of exemplars. A schema, in Kant’s parlance, is a third kind of representation which can mediate between an intuition and a concept and can thus enable the application of a concept to a given intuition. Analogously, the schematism is the process of applying the concepts of the understanding to appearances. Kant distinguished between schemata for empirical concepts like “chair” and schemata for pure sensory concepts from geometry like “triangle”, which is Buckner’s example. In Kant’s theory, which can only be hinted at, schemata result as an effect of the imagination’s task to produce a given concept’s image (KrV, AA: A140/B179). While the understanding produces a concept that acts like a rule, the imagination produces a general *Gestalt* (the schema of that concept) and a concrete image, either in free association or based on sensory input. Schemata are sensory by being imagistic representations, and yet general rather than merely particular, guided by the rule provided by the concept. I will leave it at these brief remarks on schemata, as there are many interpretative problems to do with this important notion (see Pippin 1976, Pendlebury 1995 and Matherne 2014 for further discussion). The upshot is, of course, that this consideration leads us away from a purely empiricist account towards a mixed account that incorporates rationalist elements. It is an area of research that deserves a closer look on another occasion.

Another way in which Kant’s theory of cognition is quite different from the typical empiricist approaches has to do with the central notion of spontaneity. As Evans notes, apart from the passive sensibility which only receives information, all other faculties involved in cognition in the narrow sense contain a spontaneous element. It is crucial for Evans’ interpretation that spontaneity is free of any constraints, such that the cognitive agent is “continually constructing the program” that she can execute, being “free to construct *any rules whatsoever* – as long as they satisfy the unity conditions” (Evans 2021). Yet, Evans is well aware that a number of features that are important for Kant are either not represented in the Apperception Engine or represented differently, for example, space, time, and

---

20 In addition to interpretative difficulties, it may be problematic that Kant did not rewrite the chapter on schematism from the first to the second edition, although many of his reconceptualizations, including that of the place of the imagination, might have effects on this overall view.
self-consciousness. Thus, while Kant takes the spontaneity of the understanding to be typically a self-conscious activity – i.e., an activity being conscious of itself (B 153) which allows the subject performing the spontaneous synthesis to become conscious of “the identity of the consciousness in [. . .] conjoined [. . .] representations (B 133) –, this has as yet not been implemented in Evans’ program. Evans himself acknowledges that there is still much more work to be done in order for the Apperception Engine to be fair to Kant’s original theory. But what’s more important for his practical purposes is whether the resulting performance of the program is in need of further elements. It is of course an empirical question of whether Evans’ Kantian machine-learning approach is superior to competing deep-learning architectures or whether alternative routes are sufficient or yield better results. But given the purpose of this paper, which was to present an overview of where Kant’s conception of cognition has been and still is influential, it is fascinating to see that many of his ideas are still very much alive and relevant.

4 Conclusion

In this paper I presented a survey intended to outline various approaches developed by researchers in cognitive science and artificial intelligence with an eye on the influence of Kant’s theory of cognition on these respective approaches. As it turned out, many elements of Kant’s philosophy have been influential and are still relevant in the search for the right paradigm to explain and experimentally approach cognition. It was not the purpose of this paper to remain faithful to all of Kant’s texts and engage in Kant-exegesis. Rather, this text was an exercise to see what cognitive scientists and contemporary philosophers of mind can take from Kant’s philosophy and apply it usefully to address open questions such as what’s needed for an artificial system to make sense of sensory input, or to develop cognitive models such as the predictive processing paradigm to capture neural processing in the brain. This is a fascinating area of study. Time will tell what elements of Kant’s philosophy of mind will remain fruitful and necessary in the best theories of cognition and the development of artificial systems exhibiting general intelligence.  

21 I am grateful for funding from the Volkswagen-Foundation of my Lichtenberg-Professorship and the associated research project on Situated Cognition. Thanks also for helpful feedback and comments on an earlier draft by Cameron Buckner, Richard Evans and Dieter Schönecker and my team members Krzysztof Dolega, Marco Facchin, Paola Gega, François Kammerer, Firuze Mullaoglu, Nina Poth, Bartosz Radomski, Caroline Stankozi, Tobias Starzak, and Elmarie Venter.
References

All references to Kant’s works follow the German original, Gesammelte Schriften, ed. by the Königliche Preussische Akademie der Wissenschaften, 29 Volumes. De Gruyter et al. 1902. Where available, English citations follow the Cambridge Edition of the Work of Immanuel Kant in Translation.

GMS Grundlegung zur Metaphysik der Sitten/Groundwork of the Metaphysics of Morals
KrV Kritik der reinen Vernunft/Critique of Pure Reason
KU Kritik der Urteilskraft/Critique of the Power of Judgement
V-Met Metaphysik-Vorlesungen (Lectures on Metaphysics)


Calvo, Paco/Gagliano, Monica/Souza, Gustavo/Trewavas, Anthony (2020): “Plants are intelligent, and here is how”. In: Annals of Botany 125. Nr. 1, pp. 11–28.


Nunez, Tyke (forthcoming): “Kant on plants: self-activity, representations and the analogy with life”. In: Philosopher’s Imprint.


Schlicht, Tobias (2016): “Kant and the problem of consciousness”. In: Stephen Leach/James Tartaglia (Eds.) Consciousness and the great Philosophers.


Tolley, Clinton (2020): “Kant on the place of cognition in the progression of our representations”. In: Synthese 197. pp. 3215–3244. DOI: 10.1007/s11229-017-1625-3.


