AI & SOCIETY (2023) 38:67–87
https://doi.org/10.1007/s00146-021-01287-w

ORIGINAL ARTICLE

Empiricism in the foundations of cognition

Timothy Childers[1] · **Juraj Hvorecký**[1] · **Ondrej Majer**[1]

Received: 11 February 2020 / Accepted: 10 September 2021 / Published online: 22 October 2021

Abstract
This paper traces the empiricist program from early debates between nativism and behaviorism within philosophy, through debates about early connectionist approaches within the cognitive sciences, and up to their recent iterations within the domain of deep learning. We demonstrate how current debates on the nature of cognition via deep network architecture echo some of the core issues from the Chomsky/Quine debate and investigate the strength of support offered by these various lines of research to the empiricist standpoint. Referencing literature from both computer science and philosophy, we conclude that the current state of deep learning does not offer strong encouragement to the empiricist side despite some arguments to the contrary.

Keywords Behaviorism · Empiricism · Connectionism · Deep learning · Adversarial attacks

1 Introduction

We aim to trace the strength of support for empiricism in several debates about the nature of human cognition since the 1950s till the present. We address early behaviorist approaches to learning, connectionism, and some influential versions of deep learning in turn. Each of these approaches has been criticized for its important limitations and we will demonstrate that these limitations also undermine the assumed support for empiricism. Surprisingly, some of the argumentative strategies used to attack these different research programs are very similar. While all three approaches have been used to defend empiricist theories in cognition, we find that such usage is largely unsupported.

The long-standing empiricist tradition is based on a firm belief that knowledge and the content of the mind arise primarily, if not exclusively, from sensory input. Its paradigmatic slogan, *nihil est in intellectu quod non sit prius in* sensu, was updated in the first decades of the twentieth century to construe a picture of the world from pure experience (the phenomenalism of Russell 1914 and Carnap 1928/1967). Through a series of transformations, the tradition went from explaining internal model of the world as a collection of theories consistent with a very basic notion of observation to its various current iterations that claim the sufficient amount of data and immense processing power of deep learning networks can by themselves arrive at and possibly go beyond human-level cognitive capacities.

Alongside these philosophical developments, science also adopted empiricism early on as its fundamental approach to cognition. With the publications of Skinner and colleagues in the 1930s, empiricism became a standard scientific methodology. We stress that this methodology has strongly influenced the current domain of connectionism (Walker 1992). Moreover, it still finds its adherents in several areas of deep learning. These computer science strategies build on early behaviorism, envisioned as a radical version of empiricism. They rely on large sets of data and aim to match and overcome the achievements of human cognition just by adding sufficient computing power. Our goal is to show that all these strategies failed to deliver justification for empiricism. Instead, we will come to the conclusion that to match the human cognitive level, machine learning needs to embrace hybrid models, broadly inspired by Kantian approaches to cognition.

✉ Ondrej Majer
majer@flu.cas.cz

Timothy Childers
timothy.childers@gmail.com

Juraj Hvorecký
hvorecky@flu.cas.cz

[1] Institute of Philosophy, Czech Academy of Sciences, Prague, Czech Republic

1.1 Skinner, behaviorism, and language learning

The first step in our intellectual history is behaviorism of Skinner and his followers. We are aware that there are important predecessors to this school of thought (Walker 1992), but we do not concentrate on them as they are not as strongly theoretically founded and do not have to systematically answer challenges to their theoretical commitments. For us, the debate starts to be genuinely about empiricism at the moment when there is a serious contender that rejects empiricist assumptions. While the first part of our philosophical story relies heavily on debates about forms of language acquisition, it serves only to illustrate general foundations of associationist learning strategies and their legitimate criticism. We first introduce basic building blocks of behaviorism. Then we take up Chomsky's critique of behaviorism, beginning with his 1959 review of Skinner's *Verbal Behaviour* and continuing through the 1970s and see how Skinner and Quine altered their position in its light.

Let us start with the basic tenets of behaviorism. First, behaviorism holds that mental entities are not explanatory. This is not tantamount to a claim to eliminate the mental domain. Instead, it is meant to expel the mental as an explanatory category in psychology. Hence, the notorious 'black box' argument: whatever may go on inside the black box of our heads, i.e., a subject's mental processes, is irrelevant to explanation. Second, the radical empiricism of behaviorism is restricted to externally observable inputs (known as stimuli) and outputs (known as behavior). It subscribes to a strict empiricist understanding of what constitutes credible scientific entities. If it is to be scientific, psychology must search for correlations between stimuli and outputs. Correlations identified by researchers do not necessarily correspond with traditional psychological notions. Third, behaviorism rejects traditional (or folk) categories of psychological explanation based on thoughts, attitudes, and other psychological states. Explanations of behavior must fundamentally be based on the observed data. The work of a behaviorist should thus proceed in a piecemeal fashion. Scientists need to focus on particular sets of stimuli and on particular behaviors. The correlation of behavior with desired outcomes is at the heart of the behaviorist explanation for learning. Learning consists of links between stimulus and behavior governed by operant conditioning. In the strict empiricist fashion, it is the history of encounters with a given phenomenon that shapes up any individual future performance on the relevant task. It is shaped by the tendency to repeat behavior that is rewarded in a particular situation, and to refrain from behavior that is punished. This process gives us the principles of positive and negative reinforcement—the principles that will become crucial in artificial neural networks several decades later. The strategy envisioned by the behaviorist theory of the mind aims at a creation of an implicit list of probabilistic correlations of stimuli and responses—and little else (we will return to this point in Sect. 1.4 and later).

The behaviorist project culminates with Skinner attempting at a construction of an account for language learning in his 1957 work *Verbal Behaviour*. On his account, language is understood as behavior elicited by certain other (often linguistic) behavior. Given its highly speculative account of how we come to learn languages, Skinner's theory would probably be largely forgotten by now if it were not close to the behaviorist views held by Quine. As the criticism of Skinner transfers to Quine, one of the most influential empiricists philosophers of the twentieth century, it is worth focusing on his radical vision of the mind.

1.2 Quine and Chomsky

In philosophy, behaviorism is most famously represented by Quine. In fact, Quine and Skinner were close associates. On their picture, psychological processes are not based on internal representations or models. Instead, they take place in the purely physical space of interactions. These processes are therefore not far away from later connectionist views on categorization and other cognitive tasks, where interactions between neurons and input/output relations are all that matters.

In his work, Quine aimed to build a theory of language learning based on conditioning. When a child is presented with a red ball, for example, the child might be rewarded for uttering the word "red". The basis for learning language (and hence everything else) is ostentation, or constant pointing to objects and naming them (Quine 1950). Our primary aim is not an exegesis of Quine. Rather, we concentrate on an argumentative exchange between Quine and his prime opponent Chomsky as it is this exchange that forces Quine to abandon some of his early strict empiricist inclinations, and implicitly embrace some nativist presuppositions.

As Chomsky was recently involved in exchanges with deep learning advocates (Norvig 2017), we want to point out that a predecessor for these exchanges took place in the 1950s and afterward—namely, Chomsky's debates with Quine. Interestingly, various recently employed arguments on deep learning (see Sect. 3 below) resemble those employed in their early debate.

Chomsky questions the innocence of the notion of stimulus upon which the Quine account for language learning rests. To borrow his example, suppose we seat a subject in front of a red chair and wait for an utterance. If the subject says "red" (or "chair", or "red chair"), we praise the subject's correct utterance. Our cheering is meant to reinforce such

utterances following presentation with such stimuli. But suppose the subject says, "It smells funny in here". Then, the stimulus of the utterance must be the smell of the room (and the bombarding of the subject's olfactory receptors). In that case, what counts as stimulus depends on what the subject utters. This leads to a bigger worry. By invoking the method of ostension, one tacitly introduces an intentional vocabulary. Such vocabulary needs to specify what a speaker intends to point to through an ostensive gesture. It must also identify the link between the referent object and the co-occurring utterance. This, Chomsky (1967) argues, is nothing other than a retreat to mentalist explanation. To make sense of the response, we must refer to the stimuli. Locating the stimuli requires knowing what the response is actually *about*. The intensionality of the utterance presents a further problem. The same stimulus may invoke many responses under various modes of presentation. Without recourse to mentalism, this many-to-one relation cannot be resolved at all.

These arguments first appeared in Chomsky's review of Skinner's *Verbal Behaviour* and apply, *mutatis mutandis*, to Quine's position. They are all part of a more general problem wherein utterance and stimuli seem to be independent of one another. For instance, we often utter names when their bearers are *not* present. Broader considerations about context play a crucial role in establishing the existence of *any* relation between utterance and its referent. Quine seems to be aware of this complication and in his later writings claims that reduction of many-to-one relation can be achieved via a specific mechanism within human subjects: "…learning depends indeed on both the public currency of the observation sentences and on a preestablished harmony of people's private scales of perceptual similarity." (1995, 254). It is worthy of noticing that metaphorical language of *preestablished harmony* and *private scales* neither offer a satisfactory explanation nor present a firm empiricist stance.

1.3 Mining for sentences: the probability of utterances

That more is needed in explication of the complex relation between worldly inputs and linguistic outputs is clear to many observers. Nulty describes the situation in no uncertain terms: "The typical empirical perspective on learning the referents of single terms is that of an overwhelming problem space in which the novitiate language learner must find the correct connections for words and objects from a practically infinite number of possible couplings" (2005, 377).

Empiricists have often tried to save themselves from falling into the infinity abyss by invoking dispositional accounts, linked to probability theory. Chomsky quotes Quine's characterization of language as a "complex of present dispositions to verbal behavior, in which speakers of the same language have perforce come to resemble one another" (Chomsky 1968, p. 57, quoting Quine). He then notes that we can treat dispositions as probabilities: "[p] resumably, a complex of dispositions is representable as a set of probabilities for utterances (responses) in certain definable circumstances or situations" (ibid.).

This follows Skinner's approach, in which "the probability that a verbal response of given form will occur at a given time is the basic datum to be predicted and controlled" (Skinner 1957, p. 27). "The response *Quiet!* is reinforced through the reduction of an aversive condition, and we can increase the probability of its occurrence by creating such a condition that is, by making a noise" (ibid., p. 35). All of this greatly depends on the notion of resemblance[1] of the context for utterance, something Skinner spends a great deal of time with (as does Quine 1969a, b).

The aim is ultimately to eliminate intensional idioms and replace them with probabilities determined by frequencies. Skinner invents the world 'tact' to describe spontaneous behavior in the presence of non-verbal stimuli (such as the presence of a dog, which could prompt the response, "Oh look, a *doggy*!"). In his discussion, he observes:

> [i]t may be tempting to say that in a tact the response "refers to," "mentions," "announces," "talks about," "names," "denotes," or "describes" its stimulus. But the essential relation between response and controlling stimulus is precisely the same as in echoic, textual, and intraverbal behavior. We are not likely to say that the intraverbal stimulus is "referred to" by all the responses it evokes, or that an echoic or textual response "mentions" or "describes" its controlling variable. The only useful functional relation is expressed in the statement that the presence of a given stimulus *raises the probability* of occurrence of a given form of response. (ibid., p. 82)

[1] Two quotes, one from Skinner, the other from Quine, indicate that they share their strategies at this point. First Skinner: "The probability of emission of a response is greatest when the stimulating conditions closely resemble those which have previously prevailed before reinforcement. But past and present circumstances need not be identical; indeed, any aspect or feature of the present situation which resembles the situation at the time of reinforcement may be supposed to make some contribution to the probability of response" (1957, p. 46). Quine concurs: "The situations that command assent to a given observation sentence will not be quite alike. They will be similar by our lights and by the lights of other speakers. But we can count on a curious tolerance of spatial reorientation in these similarity standards." (Quine 1978, p. 158). Notice that Skinner does not speak of relevant resemblances. On literal reading of his quote, any similarity will do. Given general philosophical difficulties with the notion of resemblance, his argument is all-encompassing and thus empty. Quine at least gestures toward *curious tolerance* in a detection of similarity standards, but this is far from a satisfactory explication.

This quote is a clear illustration of the attempt to reduce intensional idioms to an extensional notion (counting the frequency of occurrence). But the easiness with which behaviorists move from referential relations to purely statistical ones, occurring in-between utterances, remains very problematic. Quine's radical metaphysical physicalism makes things even worse. For him, physicalism means that, ultimately, there is nothing in the universe other than atoms moving in the void. Given this underlying approach, behaviorism is a natural fit, as stimuli and responses are observable parts of the natural world, unlike mental entities. It follows that, fundamentally, the probabilities mentioned above are not linking stimuli with sentences, but only one set of physical events with another.[2] In Sect. 2.4 we will demonstrate how analogical strategies based on statistics are used in machine learning to achieve human-level cognitive capacities.

1.4 The poverty of stimulus (via probabilities of utterances)

Chomsky attacks Quine's behaviorism by arguing that the frequency of an utterance following some occurrence is effectively zero:

> "...assuming 'circumstances' and 'situations' to be defined in terms of objective criteria, as Quine insists, it is surely the case that almost all entries in the situation-response matrix are null. That is, in any objectively definable situation, the probability of my producing any given sentence of English is zero, if probabilities are assessed on empirical grounds" (Chomsky 1975, pp. 310–311).

Hence, he concludes, the probability of producing a sentence in Japanese is the same as the probability of producing a sentence in English (i.e., zero) (Chomsky 1975, p. 311). This consequence follows from Quinean naturalism, as utterances in various languages are just physical phenomena and as such cannot be distinguished from each other on any non-physical ground. This makes statistical analysis a very unsuitable starting point for the endeavour of linking utterances in various languages to their referents or causes. The discussion presented here is a probabilistic restatement of the *poverty of stimulus* argument.[3] This argument standardly states that children do not learn language by stimulus, response, and reward only as these elements are not sufficiently structured to fix correct utterances (of words and sentences). Put differently, children acquire language far too quickly for acquisition to be a matter of finding a proper probabilistic matrix of stimuli and responses.[4]

Interestingly, Quine does not respond by denying that a disposition towards particular types of linguistic behavior should be qualified probabilistically. He instead argues that Chomsky is focused on the wrong probabilities because the probability of a disposition toward uttering a particular sentence is conditioned by very specific circumstances—and hence, not zero at all:

> I am puzzled by how quickly he [Chomsky] turns his back on the crucial phrase "in certain definable 'circumstances.'" Solubility in water would be a pretty idle disposition if defined in terms of the absolute probability of dissolving, without reference to the circumstance of being in water. … Verbal dispositions would be pretty idle if defined in terms of the absolute probability of utterance out of the blue. I, among others, have talked mainly of verbal dispositions in a very specific circumstance: a questionnaire circumstance, the circumstance of being offered a sentence for assent or dissent or indecision or bizarreness reaction. (Quine 1972, pp. 444–445).

Similar attack is waged by MacCorquodale:

> Chomsky seems not to grasp the difference between the overall probability of occurrence of an item in a speaker's verbal repertoire, which is the frequency with which it occurs in his speech over time without regard to his momentary circumstances, and the momentary probability of a given response in some specified set of circumstances. (See, for example, Chomsky, 1959, p. 34). The two probabilities are very different. The overall probability that any speaker will say, for example, 'mulct', is very low; it occurs rarely in comparison with such responses as 'the' or 'of. The probability that he will say 'mulct' may become momentarily extremely high, as when he sees the printed word. Of the two, overall probability is a typically linguistic concern, while momentary probability shifts are, in a sense, the very heart of the psychologists' problem, since they reflect the relation between speech and its controlling

[2] For Quine's radical physicalism, see the following quote: "I am a physical object sitting in a physical world. Some of the forces of this physical world impinge on my surface. Light rays strike my retinas; molecules bombard my eardrums and fingertips. I strike back, emanating concentric air waves. These waves take the form of a torrent of discourse about tables, people, molecules, light rays, retinas, air waves, prime numbers, infinite classes, joy and sorrow, good and evil." (Quine 1976, p. 228).

[3] We note that despite some criticism, this argument is nonetheless accepted by the overwhelming majority of linguists (for a survey, see Cowie 2017).

[4] One of the reviewers had pointed out to us that we are neglecting the central role of the context in Quine's picture of language learning. While aware of this deficiency, we believe that an introduction of such a notion only adds a further computational complexity to the already complex schema. Instead of facilitating an establishment of a connection between a term and its referent, the introduction of further variables, given by context, make the situation close to intractable.

> variables. Under what conditions does an organism speak an item from his repertoire? Simply knowing the repertoire tells us precisely nothing about that (MacCorquodale, 1970, p. 88).

Before we comment on the general issue, let us briefly comment on two problems we see in MacCorquodale's critique. His notion of *momentary probability* is utterly idiosyncratic and does not refer to anything in a regular literature on the topic. More importantly, his claim that under certain circumstances the relevant probability becomes "extremely high" is unwarranted, unless we already know how the language functions. Yet for knowing more about the functions of language, one needs to invoke the vocabulary of intentions, reference, ostentation, circumstances and other related phenomena that make behaviorist reading unlikely. If, as we noted before, the probability ultimately links two physical events, there is no good reason to assume immense fluctuations in probability distribution.

Overall, if conditional probabilities are to offer a solution, we need to consider how they are obtained. Obtaining conditional probability through counting occurrences of pairs of words, phrases, and sentences would require immense amounts of data.[5] Given that this requirement is unrealistic in the case of human cognitive functioning, it seems some additional apparatus is needed to provide the relevant conditional probabilities. But reference to any such apparatus would, as Chomsky charged, also seem to reintroduce a priori notions and thereby weaken the empiricist inspirations of the Quinean project. As we will see, this debate on prior conditions for processing closely resembles debates on setting parameters within deep learning networks, which is addressed in the final parts (Sect. 4.2) of the paper.

Alongside the above-mentioned issues, additional problems appear within the behaviorist paradigm. One of the most critical involves the notion of abstraction. Humans can abstract from concrete particulars to categories and then apply these categories to other exemplars (we might move from a golf ball, say, to the notion of a 'sphere' then on to a ball ornament for a Christmas tree). Abstraction must employ mechanisms other than a basic form of induction over instances. Indeed, it is hard to see how any form of induction could lead to a formulation of claims about abstract objects. This is yet another restatement of the poverty of stimulus argument—no amount of exposure to stimuli gets us to abstract notions. Quine, of course, is aware of this difficulty. He thus supplements his account for language learning with the conditional notion of 'analogical synthesis' (for discussion, see Gibson 1987). Chomsky responds that this notion is empirically empty; it can only serve as a serious fix to the problem of abstraction if we are provided with an exact account of what it is and how it is used (1969, p. 56). Section 3.2 will recount how the problematic issue of abstraction has resurfaced, decades later, within the debate on deep learning.

Under the argumentative pressure, Quine later appears to change his position. He denies that the introduction of mechanisms other than the correlation of stimulus and response amounts to an abandonment of behaviorism. For him, "empiricism of this modern sort, or behaviorism broadly so called, comes of the old empiricism by a drastic externalisation. The old empiricist looked inward upon his ideas; the new empiricist looks outward upon the social institution of language…. Externalised empiricism or behaviorism sees nothing uncongenial in the appeal to innate dispositions to overt behavior, innate readiness for language learning" (1969a, p. 58). In another paper from the same year, he is more specific:

> Language aptitude is innate; language learning, on the other hand, in which that aptitude is put to work, turns on intersubjectively observable features of human behavior and its environing circumstances, there being no innate language and no telepathy. … Chomsky says 'I postulate a pre-linguistic (and presumably innate) 'quality space' with a built-in distance measure'. But 'postulate' is an odd word for it, since a quality space is so obviously a prerequisite of learning, and since distances in a quality space can be compared experimentally. Quine (1969b, p. 306).

Yet the notion of the linguistic aptitude remains unexplained and unless we learn more, its collapse toward Chomskian model of the mind is likely. In any case, this is indeed a very different empiricism than the traditional one we started from.[6] We have moved from the original Humean notions where nothing is said about the content of the mind to the Kantian picture with inner mechanisms and structures that shape up incoming stimuli. It marks the beginning of a tendency of hybridization that we will observe repeatedly. Hybridization is a process of enriching the originally pure empiricist position with nativist or explicit representational elements in the face of mounting criticism. While hybridization is not a problem in itself, it makes the defense of the

[5] Many current machine learning algorithms for language translation use the tactic of counting co-occurrences of individual words in immense data sets. Seminal work was conducted by Och in his (2002) dissertation. Yet the question remains whether such a strategy is employed by humans, whose behavior Quine was attempting to explain.

[6] See, for example, Quine (1976, 57): "…the behaviorist is knowingly and cheerfully up to his neck in innate mechanisms of learning-readiness. The very reinforcement and extinction of responses, so central to behaviorism, depends on prior inequalities in the subject's qualitative spacing, so to speak, of stimulations… Innate biases and dispositions are the cornerstone of behaviorism, and have been studied by behaviorists.".

original empiricist position significantly less compelling.[7] This argument will come to the fore in the Sect. 3 with the discussion of the resource inefficiency of deep learning.

1.5 Beyond behaviorism

While essential in taking scientific psychology off the ground, behaviorism, with its extreme empiricist tendencies, became unattractive. It did not deliver on its original promises to provide a firm background for reliable connections between stimuli and responses. Given the computational difficulties of such a task, behaviorists faced ever increasing problems with restraining the domain over which probabilistic operations link inputs and outputs. More importantly, Chomsky has forced empiricists to acknowledge that even if such a domain exists, it has to take into account inner mechanisms of the mind. This is an important sign of demise of behaviorist original aspirations. Instead of a pristine black box, we now have a black box with functions, parameters and inner structures. Once these are postulated, it is only a matter of time till we learn more about their precise values and mutual dependencies. The result is a model of the mind, where it would be odd not to label its inner workings as mental. Once this is allowed, the behaviorist program is over. A parallel story can be told about empiricism. As soon as Quine moves from input-based dependencies to amend his approach with "innate dispositions" (see Sect. 1.5), he is significantly weakening its status. One might still insist (as he does) that we are just altering empiricism. After all, ever since Kant, internal constraints on the mind's processing of external inputs have been widely acknowledged (Van Cleve 1999). The question to be answered is how such constraints are to be mapped out for empiricism to deserve its name. As the reader will see in the end of this paper, there is growing evidence that chances to preserve a credible version of empiricism are rather bleak.

1.6 Computationalism as anti-empiricist movement

Partly due to its inability to respond to the Chomskian criticism and partly due to unavailability of viable alternatives, two decades following the 1960s have seen an almost complete rejection of empiricist approaches. Computationalism has ruled supreme, with its emphasis on explicit inner representations that are both installed in early computers and postulated in biological minds.[8] Computationalism equates the mind with a device that employs algorithms on symbols endowed with meaning. Its starting point is psychological. It takes for granted traditional psychological mental states (beliefs, desires etc.) and asks which processes are such that allow for these states to process worldly inputs and culminate in appropriate actions. Its view of the mind is compositional and it takes concepts to be the basic building blocks of cognition (Pylyshyn and Demopoulos 1986). Concepts are combined in accordance with algorithmic rules to create novel contents that can be cognitively utilized in various ways (stored in memory, used in arguments, acted upon, etc.). There are several persistent difficulties with computationalism. We have little idea how tenets of computationalism relate to brain's neural underpinnings of cognition, how concepts get their meanings or how syntactic brain operations preserve semantic mental relations. Yet these are not the topics to be addressed in our paper. We only want to contrast the explicitly representationalist story of computationalism with both that of its behaviorist predecessor and its current successors in machine learning. In a stark contrast to both of them, defenders of computationalism declare a strongly anti-empiricist stance, with concepts as representations being the precondition of processing of empirical inputs and delivering relevant outputs. Where do concepts come from is an uneasy question, yet some version of nativism seems acceptable for most in the field (Fodor 1998; Marcus 2018a). Given various successes of computationalism across many domains, it looked for quite some time that empiricism is a completely untenable position. Yet with the onset of connectionism in the 1980s, nativist assumptions have gradually lost their upper hand.

2 Connectionism

Out of assumptions about an alternative model of the mind, and also in response to the problems with computationalism briefly outlined in the previous section, a novel approach emerged in the late 1980s in the form of *connectionism*. It is important to note that main terms and ideas that eventually paved the way toward connectionism still play a substantial role in machine learning today. Our task is not to trace developments within connectionism in any details (for a very detailed exposition, see Schmidthuber 2015). Rather, we concentrate on those strategies within connectionism and its current iterations that tentatively support empiricism. We

[7] Gary Kemp (2006) argues that Quine and Chomsky do not, in fact, much disagree. Kemp (2006) further notes Quine could reject empiricism (read as a stimulus/response account of language learning) on the basis of his naturalism (pp. 165–7, for example). But Chomsky sees Quine as changing his views (Chomsky 1979, p. 85). We think it more likely that Chomsky is correct. Otherwise, it is difficult to explain Quine's stress on the continuum from animal learning to infant learning (for example, see Quine 1975: The Nature of Natural Knowledge).

[8] For the purposes of our argumentation, it makes no difference whether the representations are innate (nativism) or obtained by some other mechanism. It is their presence that calls for qualification of the empiricist theory.

thereby pay closer attention to the early installments of connectionism because we believe many of the current worries about empiricism in the domain of model-free machine learning and deep networks (that we discuss in Sects. 3) can be traced to its founding principles in Rumelhart and McClelland (1987).

The major appeal of connectionism lies in its offer of a parsimonious model of the mind. Unlike a complex computer with a library of concepts and a set of commands, the connectionist model consists of a simple network made up of three (or more) fundamental layers. On one end of the network, there is an input layer that roughly corresponds to sensory apparatus. On the other end of the network, there is an output layer that stands for an action. This latter layer often consists of nothing other than a binary node with a Yes or No indicator. The most important part lies in between the input and output layers: the so-called hidden layers. This is a complex web of nodes, often structured into several sublayers, where all of the substantive computation takes place. In its basic architecture, each node of the inner layer is usually connected to all of other nodes. The inner layer computes its output by summing up inputs from incoming connections. If the weighted sum passes a certain threshold, the node sends a signal further and strengthens its connectivity with successive nodes. Conversely, if the weighted sum does not pass the node's threshold, connectivity is weakened. This process of gradual change in the connection strength between the nodes aims at optimizing correct outputs for given inputs for the system as a whole. Given what has been said, neural networks training can be understood as providing a transition function between inputs and outputs. The function optimization is obtained by training the net on considerable quantities of data.

2.1 The appeal of connectionism

It is indeed remarkable that these simple networks, often containing just a few dozen nodes and a corresponding number of connections, have achieved significant results across various fields of cognition. Three elements of connectionism are worthy of special attention. Connectionist networks are very simple, with no explicit conceptual structure of their own. They are also modelled on the brain's wiring, thereby overcoming an objection of the disconnection of computationalism from its underlying biological substrate. Their flexibility allows for the experimental testing of various models of cognitive processing. Finally, in their thorough associationism, they are as close to an implementation of empiricist ideas as possible.[9] In diverse categorization and recognition tasks, artificial neural networks have surpassed expectations and performed very —often approximating or even outperforming human subjects. While remarkable in their own right, their success has led their defenders to rethink the basic elements of cognitive mechanisms (for a more detailed exposition, see Rumelhart 1998). As networks achieve their results primarily by being trained on large amount of input data, one revolutionary consequence of their success is implicit support for empiricism.

Advocates of connectionism stress that networks are modelled on the actual (albeit extremely simplified) biology of the brain. The functionality of both neurons and network nodes depends on their interconnectivity and processing of incoming signals. Points of similarity do not end at the level of architecture. They are also manifest at the cognitive level: like human minds, neural networks are well equipped to work with degraded, context dependent, and multivariate inputs. Unlike in classical computational models where the damage of a single component often leads to devastating consequences for the overall task performance, the failure of one or even several network nodes is hardly noticeable. The output might be slightly more inaccurate, yet it remains reliable. Networks also master situations where the damage occurs not within their architecture, but with the stimulus. Just as human beings can perceive and recognize objects under visually challenging conditions, connectionist networks produce correct outputs for partially occluded, degraded, noisy, or otherwise altered stimuli.

Despite their operational successes, there is a limited way to ascribe individual features or recognizably human categories to a particular node.[10] Representations of features relevant to performance of a given categorization task are widely distributed. This has been an exciting piece of news to all defenders of anti-representationalism because it seems to indicate that the artificial systems have no need for robust, semantically evaluated representations. If they prove successful across a variety of tasks, networks might shed a light on human minds as well. Maybe our minds are also devoid of representations and the contrary assumptions have been held due to our inability to shed off folk psychological intuitions.

[9] As Walker (1992) has clearly demonstrated, crucial components of connectionism were firmly in place within psychological and neurological literature long before the onset of the research program. The essential success of connectionism lies in its ability to put them all together and implement them in an artificial system.

[10] With their increased complexity, some networks now include additional constituted parts endowed with representative content. Gated neural networks, introduced by Cho et al. (2014), are example of such an approach. We claim that the emergence of similar hybrid models abandons the early-stage pure connectionist program and thus vacates the field of empiricism. Similar attempts thereby fall outside of the scope of our investigation.

Behind debates on representationalism lies a deeper question about the smallest building blocks of complex mental processes. While computationalism is fully committed to the existence of explicit representations (and has to face the worries introduced by their presence), connectionism follows in the footsteps of Skinner and Quine to exclude explicit representational features from the mind.

While the debate on the nature of representations within networks would lead us away from our main target, we want to at least indicate what is at stake and what positions were adopted by researchers on this issue. In classical computationalism, all representations are fully explicit and, whether they are taken as innate or implemented, there is no issue about their identity and specific role. Conversely, neural networks lack explicit representations, and the debate has emerged on how to account for their absence. Some authors (Rougier 2009, O'Brien and Opie 2009) argue that given the causal role representations play in any explanation of cognitive processes, it is reasonable to assume there *are* representations within the networks, albeit in an implicit form. Other authors argue that such an assumption rests on a confusion. Darwiche (2018) puts this point succinctly:

> Architecting the structure of a neural network is 'function engineering' not 'representation learning,' particularly since the structure is penalized and rewarded by virtue of its conformity with input-output pairs. The outcome of function engineering amounts to restricting the class of functions that can be learned using parameter estimation techniques…. The practice of representation learning is then an exercise of identifying the classes of functions that are suitable for certain tasks.

Whatever the outcome of this debate, it is important to see how the influence of connectionism extends beyond the field of computer science and changes the landscape within cognitive science, psychology and philosophy. While it does not promote the radical behaviorist 'black box' approach to the mind that recognizes knowledge of inputs and outputs only and ignores any of its inner workings, it makes the mind significantly less transparent than its opponents would like to have it. Were the connectionist program successful, Occam's razor would make postulations of true representations at the psychological level of explanation unlikely. Psychological terms and their adoption at the level of scientific mentalism would have to go, and the doors of empiricism would remain wide open.

2.2 Limitations of connectionist networks

While remarkably successful, connectionist networks are also prone to problems. We will only briefly mention some of them as the real target of our endeavor are analogous problems in the newest incarnation of networks within the deep learning paradigm. From the onset of their application on cognitive tasks, networks have been under attack on their incapacity to solve their target tasks properly. Notorious are exchanges between defenders and opponents of networks on the topics of compositionality and systematicity in natural languages (Fodor and Pylyshyn 1988; Aizawa 2003).

Even with a space for skepticism regarding the notion of systematicity (cf. Johnson 2004), troubles modelling compositional and systemic linguistic intuitions in connectionist networks seem to stem from the absence of any basic building blocks that could be combined to form larger linguistic units.

While we cannot delve into these discussions more deeply, we would at least like to connect them to some other issues that have been identified within the connectionist's domain. These include difficulties with generalization from concrete examples to a highly abstract level, issue with general rules or output invariance (see Mayor et al. 2014 for an overview).

All these difficulties seem to originate from an overemphasis on empiricism. Any move from the concrete to an abstract level has been difficult for empiricists since early on and the technological advances do not come to a cheap rescue.

We are not arguing that connectionist networks are a priori unable to deal with the set of problematic cases that were used as a weapon against them early on. On the contrary, many ingenious solutions to these challenges have been offered and some are in use till now. For the thesis of the paper, the problem lies elsewhere. While various solutions might have worked, they achieved satisfactory results only because the novel architectures introduced dedicated modules and specific structures within the networks to deal with various mentioned difficulties. Their introduction meant abandonment of the original empiricist credential. Once auxiliary dedicated structural features were in, the tangibility of the claim that networks operate purely on the inputs and arrive at abstractions on their own lost its credibility. While these amalgamated results of newer networks satisfied the majority of connectionist community, because the emphasis has focused on achieving the target efficiency, empiricist undertone has been lost. None argues that complex connectionist networks with memory slots and explicit categorization modules offer a strong support to empiricism.

2.3 From connectionism to deep learning

While the connectionist program has continued mostly uninterrupted, its philosophical significance has temporarily ceased. This was largely because the debate over the nature of connectionism has moved from philosophical quarrels about the building blocks of cognition into practical concerns about optimal architecture for various tasks that

systems were trained for. Yet, very recent developments in the field (since around 2015) have once again galvanized disputes about suitable approaches to cognition based on complex neural networks. The last decade has seen a revolution in artificial intelligence (AI) based on a variety of sophisticated network architectures that are often grouped together under the rubric of 'deep learning' (DL). The term usually refers to very large-scale networks consisting of tens of thousands of nodes with a multitude of mutually connected layers and additional dedicated features. While starting off from basic connectionist principles, deep learning resulted from a skillful combination of important architectural insights and technological advances that moved the field significantly forward.

Deep networks entered public imagination by proving remarkably successful compared to all other competitors. They achieved superhuman results in a variety of domains, such as image recognition and highly sophisticated game playing (checkers, chess, Go, and Starcraft), and very good results in a number of other fields (a list can be found at https://deepindex.org/). Experts recognise that without hardware advances, the field could not have achieved these successes. Geffner (2018) admits that: "The recent successes have to do with the gains in computational power and the ability to use deeper nets on more data." (p.2) In his extensive historical review, Schmidhuber (2015) describes how techniques essential for the current string of deep network accomplishments are founded on principles and theories that have been known for decades. Similar observations have been made by Darwiche (2018) who credits Oren Etzioni (see ibid, ft. 9) with the thought about not-so-novel theoretical foundations of deep learning. Whether hardware powers bear crucial responsibility for the DL successes is debatable, yet it is likely that a brute force is not solely responsible for its successes. Without specific architectural advances that we comment on in the next section, the field would not achieve what it did. Darwiche (ibid.) concurrently speaks of the employment of better statistical methods for data fitting in various current approaches. Skeptically, he also assigns a role for the hype surrounding the deep learning to the downgrading of the measurement of success. For example, in the domain of language translation, early challenges in machine translation quantified the success rate of a system to translate a previously unknown text to a foreign language and back while novel approaches are fine with giving us a reasonable enough estimate of what the target text in foreign language is about. The so-called *gist translation* would fail the early criteria on translation success miserably, though it became the recent new standard of achievement.

We are not in a position to adjudicate a debate about the causes for DL successes. Instead, our investigation is interested in its philosophical significance. Despite the notorious triumphs of DL and vocal voices to the contrary, we believe there are good reasons for skepticism about the kind of support deep networks provide for empiricism. Our aim in this part of the paper is twofold. While we want to show that current debates on the nature of cognition via deep network architecture echo some of the core issues from the Chomsky/Quine debate, our purpose is to go further. We aspire to provide additional arguments for skepticism with regards to the overall tangibility of empiricism.

2.4 Empiricism and deep learning

There are many types of deep networks, differing in architectures and the presence of specialized submodules. Instead of a simple forward-feed model where information flows from input nodes through internal nodes towards output ones, current more complex nets include backward loops and dedicated memory modules. Because the field of deep networks is multifaceted and the optimization tasks remain of a prominent interest to the majority of researchers, only few nets are built with intentionally empiricist principles in mind. Many researchers are aware of the distinctions between bottom-up purely data-driven networks and more complex setups that includes internal modules, various gates or encoder/decoder architecture (see Baroni 2020). All of these additional architectural features have been utilized to enhance networks' performance. As we have already indicated, our primary concern lies with the first, seemingly simpler group of networks, devoid of dedicated internal modules. There are various ways to categorize these distinct types of nets. Darwiche (2018) speaks of model-based and function-based approaches to AI, with the latter determined solely by inputs. Geffner (2018) distinguishes between solvers and learners. While the solvers compute their outputs in accordance with a model, learners are empiricist in nature, driven by the data. Geffner further divides learners into two sub-groups: deep learners and deep reinforcement learners. It is the last group that is of most interest to us as it utilizes "a non-supervised method that learns from experience, where the error function depends on the value of states and their successors" (2018, p. 2), while standard deep learning regularly relies on supervision. An analogous distinction of model-free and model-based systems is drawn in more details by Lake et al.:

> The statistical pattern recognition approach treats prediction as primary, usually in the context of a specific classification, regression, or control task. In this view, learning is about discovering features that have high-value states in common—a shared label in a classification setting or a shared value in a reinforcement learning setting—across a large, diverse set of training data. The alternative approach treats models of the world as primary, where learning is the process of model building. The difference between pattern recognition and

> model building, between prediction and explanation, is central to our view of human intelligence. (Lake et al. 2017, p. 3)

In this text we will follow the terminology of Lake et al. and speak of model-free and model-based systems, with model-free approach being our primary concern.

The quotes above might make one believe that both systems are essential for understanding cognition. However, a part of the commotion surrounding the recent advances of deep learning stems from a conviction of some of the researchers that there are good reasons to concentrate exclusively on the model-free variants while regarding model-based approaches as secondary. If so, then the combination of ever larger data sets and brute force computing brings upon the final justification of empiricism. When Silver et al. (2017) describe their famous AlphaGo Zero system, they use unequivocal language:

> … our results comprehensively demonstrate that a pure reinforcement learning approach is fully feasible, even in the most challenging of domains: it is possible to train to superhuman level, *without human examples or guidance*, given no knowledge of the domain beyond basic rules. (emphasis ours)

They then continue by contrasting the performance of their system with that of human beings:

> … humankind has accumulated Go knowledge from millions of games played over thousands of years, collectively distilled into patterns, proverbs and books. In the space of a few days, *starting tabula rasa*, AlphaGo Zero was able to rediscover much of this Go knowledge, as well as novel strategies that provide new insights into the oldest of games. (emphasis ours).

The reader should take into account straightforward gestures toward empiricism in both quotes. Some researchers even argue that model-free systems can do all the work that we expect from any cognitive system. Ng endorses a full (present!) replacement of human cognitive capacities by artificial systems: "If a typical person can do a mental task with less than one second of thought, we can probably automate it using AI either now or in the near future." (Ng 2016).

It is difficult to take these quotes seriously. While partly justified by practical successes of their systems, these quotes rely on strong philosophical assumptions which have been sufficiently scrutinized neither by the computer science community nor by the external experts. To defend empiricism in the form of the model-free approaches requires a thorough philosophical exercise. At least one philosopher has offered his methodical support for these optimistic judgments. In his 2018 paper, Buckner argues that most direct philosophically support for empiricism comes with *convolutional* deep networks. While relying on the already mentioned hardware advances, three characteristics have combined to make convolutional DNs strikingly powerful: network depth, convolution, and pooling (Buckner 2018, p. 5350). Originally, networks' depth referred to the number of its layers. The greater number of layers, the deeper the network. In accordance with this assessment, a straightforward analysis might take the early connectionist attempts as fairly shallow and recent developments as significantly deep. Yet, the current complexity of the architecture might call for a more detailed assessment of depth criterion. Several suggestions from the field go beyond a simple counting of layers. Schmidhuber (2015) highlights the role of credit assignment paths in tracing the causal origins of a given output, while Sun et al. (2016) concentrate their effort on the effectiveness of margin bounds. On all approaches, the depth reflects a measure of the complexity of network processes, and proves an ever-increasing hierarchical complexity of network architecture. From the perspective of empiricism, it is important to note that the greater depth permits of a more nuanced analysis of the input data. In combination with other two mentioned features (convolution and pooling), layers at different depth can focus on distinct features of the input. The invention of convolution altered the foundations of neural networks substantially. Convolution refers to filtering algorithms that select certain input features as belonging to a particular category. For example, filters may discern the presence of an edge or a colour at a particular location in a visual object recognition task. By doing so, convolution extracts from the image specific features that lead to its ultimate classification. Crucially, the combination of depth and convolution allows for detection of invariances that are common across various target objects from a given category but might not be straightforwardly detectable at each instance of object presentation. For example, in detection of faces, eyes might be once fully visible, once seen only partially from an angle and once completely hidden, yet the network learns that despite these noticeable differences, one category of objects (a face) is always present. Finally, the pooling mechanism assesses detected lower-level features and sends an affirmation of their presence to a next layer. It works by averaging over the results of convolution, or by using a particular down-sampling operation such as max-pooling (for details, see the cited piece by Buckner). From the perspective of empiricism, pooling allows for capture of categories across various layers of complexity. Analogously to the processing of images in brain's visual areas, networks can work their way out from the detection of the simplest features like lines and shadows all the way to the high-level categories like faces. Thanks to these three characteristics of depth, convolution and pooling, networks can learn to classify a set of objects despite a great number of nuisance variations in their presentation.

Given that convolutional networks arrive at their results solely with the help of the three above-mentioned features, they present for us a prototypical example of function-based, model-free architecture. As such, they bring in the approach that is the closest to the empiricist mind, thereby deserving a detailed critical scrutiny. In the remaining part we will demonstrate that model-free networks of this type face their own serious limitations, significantly weakening support for the thesis that artificial minds can be empiricist.

3 Limitations of deep learning

We now dive deeper into problems with specified types of the model-free networks. Our starting points find inspirations in Garnelo and Shanahan (2019), who introduced several avenues for criticism of deep learning. They locate three types of deficiencies[11] with deep learning: (1) data inefficiency, as DL requires vast amounts of data; (2) poor generalization; and (3) lack of interpretability, as deep learning achieves its classificatory task along a pathway that we have difficulties to interpret. We are about to analyze all three of them and focus on their role in the undermining of DLs empiricism. We also add one more problem, that of transferability, because it is closely related to our main concern. It is also worth noting that several authors have detected other important difficulties with the overall performance and promises of deep learning (Marcus 2018b; Pearl 2018). While appreciating their contribution to the overall discussion on the feasibility of the deep network research program, we are not addressing their additional worries, because they bear little or no relation to the question of empiricism.

In the next sections, we address each of the concerns and link them to the issue of the tenability of empiricism. While the issue of resource inefficiency illustrates a mismatch between deep learning and human cognition, it is the other three concerns that display failures to support empiricism. Issues with the failure of abstraction indicate a crucial weakness in model-free architecture that we have little idea how to remedy. This weakness prevents deep networks to both get a grasp on some of the most general categories of cognition and understand causal underpinning of worldly events. Transferability failures show lack of robust representational powers on the side of the nets. These are further demonstrated in a set of opacity concerns. The opacity of deep learning algorithms usually means that "the computations carried out by successive layers rarely correspond to humanly comprehensible reasoning steps, and the intermediate vectors of activations they generate usually lack a humanly comprehensible semantics" (Garnelo and Shanahan 2019, p. 17). Yet, this partly epistemic reading of opacity is secondary to us. The real limitation, with regards to empiricism, comes from seeing how opaque nature of deep networks creates its own idiosyncratic categorization structure. The resulting categorization is then inherently vulnerable to various kinds of adversarial attacks. Upon discussing all these various difficulties, we will finally be able to spell out parallels to the Chomsky/Quine debate and address the current status of empiricism directly.

3.1 Resource inefficiency

The first category of problems arises from quantitative requirements on deep learning. Data dependence is listed among the most perplexing issues in the field (Tan et al. 2018).

Let us compare the learning curve for human beings and networks on identical tasks. The often-cited successes of deep networks in game playing seem less impressive when considering how humans can achieve similar results with much less input—and much more quickly (see Fig. 3 in Lake et al. 2017). To attain the celebrated victory in the game of Go, the winning algorithm had trained on a number of games that would correspond to some 200,000 "human" years of playing the game—relying on extremely large data sets. Networks typically demand training sessions that last for tens (and hundreds) of thousands of trials. It is only after this amount of exposure that they are able to categorize objects from a given domain. This does not even remotely match the human ability to learn from just *one* example (Lake et al. 2015).[12] It is unclear (and improbable) whether increasing the number of nodes and layers or tweaking the internal structure of a network can remedy such a huge discrepancy.[13]

While the belief in the brute force paradigm of ever-increasing computational power to overcome resource inefficiency remains popular, there is an increasing skepticism about the approach. One of the most persistent objections is the lack of learning-to-learn strategies within networks (Lake et al. 2017). This limited ability to come forward with

[11] In our terminology that better fits our purposes we speak of (1) resource inefficiency, (2) failure to implement abstraction; and (3) opacity, respectively.

[12] Defenders of neutral networks frequently respond by arguing that even if human minds were representational systems, they would still require a huge number of encounters with data. These encounters do not take place at the individual level. Instead, priors are set by their long evolutionary history. For an extensive discussion of this issue, see Lake et al. (2017, Sect. 5.1).

[13] There are some attempts to build nets that handle one shot learning, yet these cannot be easily repurposed for other purposes and still require significant prior exposure to examples within the same target category. Most often, Siamese deep networks (Bromley et al. 1993) with triplet loss function are employed for these types of tasks.

new learning strategies is what prevents model-free systems from reducing the size of data sets, required for training.

Yet, we also want to argue that, while indicating a significant gap between humans and networks, resource inefficiency is in itself not an argument against generic empiricism. If the crux of empiricist commitments relies on derivation of categories from pure data, we should not be surprised that immensely large data sets are required. On the other hand, an attempt to explain human minds in traditional empiricist terms loses its attractiveness due to the significant dissimilarity in efficiency of learning strategies between networks and humans.

3.2 Abstraction

Abstraction is a broad notion and when researchers report failures of networks to abstract, they often mean different things. The most common usage is that of obtaining invariant categorial information about a given target from its tokens. A related usage refers to a possibility to derive general rules from circumstantially distinct instances of certain phenomena. While the first notion of categorization is crucial for various discrimination tasks, the second is often invoked in language comprehension and production. Finally, there is a third layer of abstraction, that of uncovering causal relations between phenomena. It plays a dominant role in science, but is also prominent in folk theoretical explanations. We will briefly demonstrate that, within the field of model-free deep learning, processes of abstraction face difficulties at all three layers.

3.2.1 Categorial abstraction

While the concept of abstraction is not precisely delineated, several important notions seem to play a role in judging an operation as abstraction. On the most common reading, abstraction amounts to an *extrapolation* of a relevant category from concrete examples within a particular domain. Taylor et al. (2015) define abstraction as

> a process of creating general concepts or representations by emphasizing common features from specific instances, where unified concepts are derived from literal, real, concrete, or tangible concepts, observations, or first principles, often with the goal of compressing the information content of a concept or an observable event and retaining only information which is relevant for an individualized goal or action.

This process can achieve various levels of generality as any target input can be subsumed under several distinct categories. The task of a network is to learn to detect and reliably track criteria according to which targets are subsumed to the particular category. However, there is some inherent difficulty in abstraction tasks for the networks:

> In classic abstraction, states that are similar with respect to a property of interest are merged for analysis. In contrast, for NN, it is not immediately clear which neurons to merge and what similarity means. Indeed, neurons are not actually states/configurations of the system; as such, neurons, as opposed to states with values of variables, do not have inner structure. Consequently, identifying and dropping irrelevant information (part of the structure) becomes more challenging. (Ashok et al. 2020).

As the quote indicates, the very architectures of networks makes it inherently difficult to select relevant invariance features and suppress the redundant ones. To put it differently, networks have to find a way to compress information about their input while preserving only information pertinent for generalization to yet unobserved examples from the same category. Undoubtedly, it is a daunting task for any empiricist.

In line with the empiricist assumptions, it has been long assumed that networks do not search for predefined features as they only generalize from the input data and nothing else (Ramsey and Stich 1991). However, even before network training process starts, fundamentals of abstraction are smuggled into the learning process by a pre-classification of a training set. The presence of classificatory labels that are associated with training sets constitutes implicit comparative patterns, providing a springboard for abstraction. Successful image recognition, for instance, is based on images that have been pre-classified by hand. The provision of prior classifications can create the illusion that networks are classifying all by themselves. Yet without supervised pre-labeling, no classification would be possible. This supervised pre-labeling is not to be confused with a more general notion of supervised learning. While supervised labeling delineates the target category, in supervised learning the network receives feedback on the precision of its categorization processes. Even in unsupervised learning, feeding the network unlabeled content from within *one* category is sufficient to provide implicit category membership. Neglecting the crucial role played by pre-classified input generates a misleading notion about the spontaneous emergence of a necessary categorial structure. We also want to point out that the presence of implicit labels that result from pre-classification does not constitute a problem in itself. Empiricism operates with labels—after all, they are to be found everywhere and serve empiricists just like any other input. On the empiricist picture, minds are learning to label particular items, properties or events. The real problem consists in the fact that pre-classified inputs are fed to the network as if they were raw data, when in fact they are pre-processed by human minds.

Whatever the origins and hidden characteristics of the input data, the question remains whether a network is capable of real abstraction over its inputs. Buckner (2018) argues that, due to their specific architectural design, deep networks are indeed conducting a genuine process of abstraction. He speaks of a specific *transformational abstraction*, which employs all three essential building blocks for networks, mentioned in Sect. 2.4. Their depth allows an input to undergo hierarchical processing that, with each step, abstracts away from the particularities and fixates a categorial invariance. During the process, the network learns to ignore a number of nuisance variables of token inputs and to acquire relevant categories instead. Convolution functions as a filter that detects essential features and leaves aside all the others. Concurrently, the pooling operation decides about the presence of categorial features to be delegated to a higher processing layer. It does so across a larger detection area, thereby determining whether a feature is a local aberration or occupies a more significant position and is therefore crucial for the categorization. A subsequent layer conducts the process again, this time searching for a more abstract input feature. With enough depth, networks can learn to classify vastly divergent stimuli within a single high-level category while neglecting their idiosyncratic individual features.

It is easy to see why the process of transformational abstraction is seen by Buckner (2018) as justifying empiricism. While networks are building up their categories solely from inputs, it looks like we are back in the tradition of furnishing minds with experiences only. Given that Buckner explicitly ties his analysis of deep learning to the central debate on empiricism, we will address his conclusions shortly. However, one difficulty should be noted right away. Even by invoking their intricate processes, convolutional networks never arrive at some of the most general concepts, such as negation, and universal or existential quantifiers. As Buckner (2018, p. 5360) notes, "additional components … might need to be added to deep convolutional neural networks [DCNNs] for them discover mathematical or geometric properties themselves". We claim that such additional components are likely to lead us astray from the doctrine of empiricism.

3.2.2 General rules

The second layer of abstraction, that of acquiring general rules, can be illustrated by an intense battle waged since the early days of connectionism and continuing to the present. We refer to the debates (Fodor and Pylyshyn 1988; Fodor 1992) on compositionality (in vision, action, language, goal-settings, etc.). The vast complexity and productivity of cognitive processes give support to an assumption that such diversity is possible, because simpler elements of cognition are combined to create novel units. For the new units not to be random, systematic rules are needed. This overall compositional character of mental and behavioral processes brings in serious difficulties to all explanatory strategies built on empiricist foundations. In representationalist architecture, the units of combination are clearly delineated and as such they can enter as variables into place holders of general rules. In neural networks, it is not at all clear what elements could be combined in such a manner. Classical connectionism struggled with this fact, echoing Chomsky's criticisms of Quine's associationism. Unless their creators deliberately formulate a more explicit account of simpler building blocks, such as with objects in scenes (Eslami et at. 2016) or explicit subgoals in an action (Kulkarni et al. 2016), novel deep learning approaches also face difficulties accounting for this phenomenon.[14] Yet moves to enrich networks with additional dedicated structures resemble Quine's eventual abandonment of his purely empiricist strategy for a hybrid view that does not reconcile well with his original intentions.

Problems with abstracting towards general rules directly follow from an assumption that networks learn by association only. That assumption makes learning *general* rules especially troubling. But are not general rules a primary resource for our cognitive make-up? We do not necessarily have to think of complex moral rules; rules for mathematical operations or learning by induction are problematic enough. Just as one can learn to categorize images on the basis of countless examples, one should be able to obtain general rules from observing their individual instances. Yet this level of abstraction has not been observed in deep learning networks. In fact, even defenders of abstraction processes within networks acknowledge failures in this domain. For example, when Baroni (2020) analyzes an ability of networks to capture hierarchical tree structure of language, he discovers that "when … models are not provided with explicit information about conventional compositional derivations, they come up with tree structures that do not resemble those posited by linguists at all." Once again, this is not a welcoming development for a defender of model-free approaches.

3.2.3 Causality and correlation

The third level of abstraction is concerned with capturing causal relations between events. It is the very nature of association processes that they are merely uncovering correlations. Philosophy of science has taught us long ago that correlations and causations differ radically and unless networks are able to capture the latter, they will not be of

[14] We have not yet seen a decisive breakthrough (compare Lake et al. 2017, sect. 4.2.1) in the debate about how well compositionality can be captured by a connectionist network.

much help to explicate the events they observe. Throughout his career, Judea Pearl has offered many important insights on the difference between simple correlations and causality. In his recent paper (Pearl 2018), he points out inherent limitations of neural networks in tackling causal relations. Causality necessarily involves counterfactual reasoning. In causal investigation one asks whether a specified event would bring about another under different circumstances. However, associations exist between observed events only. Because networks are built as association machines, there is no place for them to handle the domain of counterfactuals. In the same paper Pearl (ibid.) points out that this does not rule out a possibility to capture causal relations by artificial intelligence. Artificial systems, enriched with various models of the world, are capable of detecting and explicating its underlying causal structure. A similar point is made by Lake et al. (2017) when he advocates for the use of explicit models in cognition: "Cognition is about using … models to understand the world, to explain what we see, to imagine what could have happened that didn't, or what could be true that isn't, and then planning actions to make it so." (ibid., 2). In this quote, we want to stress the role of counterfactual situations. The only way a network can assess for such situations is with the help of an explicit model. Yet such move cannot satisfy a defender of empiricism. Counterfactual relations can be discovered, but not by purely empiricist methods.

3.3 Transferability

The general project of artificial intelligence was originally conceived as that of building artificial systems that attain *general intelligence* or at least solve generic problems (for the very early formulation, see Newell, Shaw and Simon, 1959). While it is hard to specify what the criterion of attaining general intelligence or generic problem-solving amounts to, one requirement seems obvious. It is a system's transferability. A system can be assigned general intelligence or said to be capable of solving generic problems when its success to solve tasks in one domain can be transferred onto another domain. The scope of generality is difficult to set beforehand (should one system be able to solve quadratic equations, design reclining chairs and shoot basketball to count as a generic problem solver?), so some reasonable restrictions of the target domain is legitimate. Even with such restrictions in place, no system has demonstrated its ability to solve generic problems. However, given the input-driven architecture of model-free networks, their attainment of the general problem-solver's goal seems particularly hopeless. Their training, focused on a particular task, makes transferability of the resulting function particularly troublesome. Geffner (2018, p. 4) describes "transferring useful knowledge from instances of one domain to instances of another" for model-free networks ("learners", in his vocabulary) as particularly challenging. He also explains why that is the case: "Learners can infer the heuristic function h over all the states s of a *single* problem P in a straightforward way, but they cannot infer an heuristic function h that is valid for *all problems*. This is natural: learners can deal with new problems only if they have acquired experience on related problems." (p. 3). Importantly, it is fair to say that some transferability is achievable in deep learning networks. However, as Marcus observes, the applicability of a learnt function is surprisingly restrictive: situations "in which the deep reinforcement learning system is confronted with scenarios that differ in minor ways from the ones on which the system was trained show that deep reinforcement learning's solutions are often extremely superficial." (2018, 8) He illustrates the point on various games that networks have excelled at. Networks that mastered Atari games were very different from those successful in Go. Differences within the domain of board games (board size and its symmetry, differences between games with perfect and imperfect information) are sufficient to block an algorithm's efficacy.

Lake and Baroni (2018) make similar point in the domain of language comprehension, when they claim that recurrent networks "generalize well when the differences between training and test … are small [but] when generalization requires systematic compositional skills, RNNs fail spectacularly" (2018, 1). This is a very unwelcoming consequence of the model-free deep learning approach. Yet it is illustrative of the pitfalls of empiricism. If a system is fitted to a particular task by being endlessly trained on one kind of input data, there is only a diminished chance it would perform well on dissimilar kinds. If empiricist-based models are to bring problem-solving success, it is going to be necessarily restricted to a small subset of tasks. Transferability even within a single domain (say, board games), is virtually impossible to ensure and general problem solving remains an elusive dream.

3.4 Opacity

The property of opacity is hotly debated within the domain of deep learning networks (Burrell 2016; Zednik 2019). Opacity refers to our inability to comprehend functional dependencies within the network. We can also say that it signifies the indeterminate way networks represent humanly recognizable features of input. Due to their immense distributive complexity, deep learning networks are epistemically almost impenetrable. Tinkering with networks is often a matter of trial-by-error as we have very little idea why they perform the way they do and how they learn what they do. Consequently, in comparison to systems with explicit instructions, networks suffer from low debuggability. When

nothing is known about their inner dependencies, it is nearly impossible to fix possible errors. This observation points us once again to the 'black box' problem that we have repeatedly encountered before. While networks are not black boxes in the strict behaviorist sense, their inner workings are substantially obscured. Black boxing, it turns out, is not only epistemically significant, but it also has practical consequences.

Acknowledging epistemic and practical difficulties of opacity, we try to extend its significance to the debate on empiricism. We illustrate our point with notorious cases of adversarial attacks on deep learning networks. So far, the problem has been covered almost exclusively by the computer science community; philosophers and theoreticians of cognition have not taken a strong enough lesson from these effective efforts to block networks' performance. Adversarial attacks are simple methods of stalling the success of networks by targeting the categorization process for the very domain in which they are supposed to be expert classifiers. The unusual behavior of networks under the attacks was first noted by Szegedy et al. (2013). Since then, it has been demonstrated that the issue is more widespread than originally conceived. Adversarial interventions either alter non-robust[15] features of the target inputs (inputs still remain easily recognizable by humans) or introduce changes at the level completely imperceptible to human observers. During the attack, alterations of the input data substantially weaken or completely paralyze successful operation of a network.[16] There are two versions of adversarial attacks, white box and black box ones (Huang et al. 2017). A white box attack requires knowledge of the network's architecture. This knowledge expedites access to system vulnerabilities. While white box attacks are interesting, from the perspective of empiricism black box attacks are more intriguing. These occur with no prior knowledge of a particular network architecture. "Attackers" only know the training set and categorization task (e.g., the classification of a picture as a human face). Upon witnessing a black box attack, an uninitiated observer might be very puzzled by the behavior of the network that fails utterly, while all features of its operations remain apparently undisturbed. The fact that consolidated networks can be fooled by imperceptibly minor perturbances on the input side should cause alarm for any proponent of empiricism.

The scope of these disturbing outcomes has been placed under scrutiny by Ilyas et al. (2019). Their research team suggests a reversal of a common explanation of successful adversarial attacks. The previously held views explain vulnerability to attacks as resulting from overfitting or particular design flaws. The authors point out that non-robust features of the dataset, susceptible to the attacks, actually perform essential functions within the network. These features that humans are not capable of detecting, help to optimize the performance of the network. In an insightful experimental setup, Ilyas and his team separated humanly undetectable features within a given category (e.g., planes) that are vital for network's classification and overlapped them onto images of a different category class (e.g., animals), preserving the original labels (names of airplanes). The network was then trained on these newly conjoined images with labels from the original category set. Although the exercise was deeply confusing in human terms, the trained network was eventually able to classify a standardly labelled dataset with both images and labels that corresponded to the robust human categories. (Trained on pictures of animals that contained undetectable features of planes, together with the plane labels, the network learnt successfully to classify planes.) As the researchers concluded, "this demonstrates that adversarial perturbations can arise from flipping features in the data that are useful for classification of correct inputs (hence not being purely aberrations)" (Ilyas et al. 2019, p. 2).

There is a further worrisome consequence of adversarial attacks. As Ilyas et al. (2019) and other teams have observed, these attacks are transferable across various architectures trained on identical data sets with unexpected ease. Authors suggest why: "different classifiers trained on independent samples from [the same] distribution are likely to utilise similar non-robust features" (ibid., p. 8). Networks find humanly indiscernible input features very instrumental in their image categorization.

These recently discovered properties of network make them not only vulnerable to malicious attackers, but also inform us about principled limitations of model-free networks. Crucially, these observations of adversarial attacks should be very unwelcoming to any defender of empiricism. If the thesis of empiricism consists of utmost reliance on human experiences to obtain the cognitive content, with networks we see those experiences as we know them, are actually secondary. Some other, humanly undetectable features of input turn out to be cognitively more important than experiences.

[15] Non-robust are those properties in the input dataset that are insignificant for human categorization, yet central for categorization by NNs.

[16] Adversarial strategy can also be employed as a productive tool in cognitive operations. Gershman (2019) uses "generative adversarial algorithms" where the adversary (called generator) is a part of the system and helps the learner (discriminator) to recognize the correct samples via a specific kind of a minimax game. The generator is trying to produce samples that trick the discriminator into incorrectly classifying them as real, and the discriminator is trying to learn how to detect these fakes. This dynamic system is designed to explain certain complex psychological and neurobiological phenomena.

4 Analysis of the central problems

We have seen in the previous section that despite their successes, deep networks are engulfed in several serious difficulties. While proving their skills across diverse domains, they do not meet proclamations of some of their proponents of achieving human-level general intelligence solely from large-scale dataset training. Accomplishments of the model-free networks remain limited to specifically defined tasks. Furthermore, as the issues with adversarial attacks demonstrate, even within accomplished tasks networks are performing in a highly unorthodox manner. As we have already indicated, the above-mentioned limitations of the nets are closely tied to the empiricist nature of network's operations. Let us start with a closer look at opacity and its causes. Eventually we will expand our picture to cover all the other problems discussed above.

Together with Ilyas et al. (2019), we think the opacity emerges due to the following series of processes underlying categorization. To eventually classify a picture or an object, the network creates running hypotheses on the surveyed input and comes up with novel categorization trajectories. The optimization requirement provides no assurance that the network tracks the same set of properties that a human would use for the task. For example, in the case of pictures, the network's classificatory success apparently does not depend just on humanly discernible qualities such as colors, patches, lines, shadows, and so on. Instead, the network very likely tracks various non-robust properties that are dissimilar to anything we would expect from a human process of categorization.

This does not mean that networks *fail* to categorize a given target set. On the contrary, they often classify the set with a very high accuracy. As networks track properties that humans do not consider critical for a given task (and in some cases *could not possibly recognize* them as such, though see Zhou and Firestone 2019), there will always be a space that adversarial attacks can latch upon.[17] Moreover, altering properties that are not critical to our recognition of a given object can facilitate target recognition. These considerations lead to a question about the sameness of human and network's systems of categorization. It is highly likely that there is, at most, only a partial overlap between our categorization and those of networks.[18] Learning the function connecting inputs and outputs can be spelled out in the philosophical jargon: networks come up with a distinct category *intension*. Their trajectories of categorization form different concepts than those employed by humans.

Our interpretation of what networks achieve can be extended onto further problems we encountered in the previous section. Difficulties with the transferability of functions is a direct consequence of adopting a highly specific solution to the categorization task. Demands of optimization force networks to adopt solutions, which are strictly based on the training set. Resulting solutions are therefore inapplicable to inputs outside of an appropriately extended training set, referred to as a "training space" (Marcus 2018b). If the adopted solution is understood as intension, training space is then the category extension. After the completion of the training process, the network's intension is given by the learnt function and cannot be transferred to tasks outside of a relatively small domain. Resulting intension cannot be applied to different training spaces as these constitute distinct extensions.

Given what we have just claimed about differences in intension and extension between networks and human categorial system, we are now in a position to explicate possible sources of networks' difficulties with various types of abstraction. An ability to abstract depends on the scope of the detected features in the input. As we have seen, networks systematically detect in their inputs sets of features that are very distinct from those humans consider essential for the task. That is why we have described their resulting functions as representing different intensions. Consequently, generalizing on such a different set of features leads to an adoption of higher-order categories, necessarily distinct from our own. We allow ourselves a luxury of speculation and argue that even if nets would be able to come up with general rules and causal dependencies, they would operate on categories, distinct from those that we are familiar with.

It is time to connect what was just claimed with several of our earlier points. The marked limitations of behaviorism within the domain of philosophical psychology originally led Quine and Skinner into the blind alley of pure empiricism. There is no saving of their project without an addition of some representational features. In the present times, adversarial attacks offer an intriguing argument for the intrinsic limitations of model-free learning. The application of purely empiricist approaches to learning forces networks to adopt

[17] In the above-mentioned set of experiments, Ilyas et al. are able to eliminate non-robust features from the data set, thereby forcing a network to eventually adopt robust categories. It is, however, clear that such a move cannot be defended on empiricist ground as it artificially intervenes into inputs.

[18] Buckner hints at the interesting possibility of overlap between deep network categorizations and those of humans. Given that networks conducting a visual search look for invariants across many concrete depictions, their understanding of a visual scene resembles that of an avant-garde artist from the cubist or expressionist period, who is interested in depicting variables that do not change with object transformation.

recognitional capacities that are not picking up expected categories. Instead, networks find an intension that does not necessarily detect humanly recognizable features. To capture our categorial structure, networks would have to be fitted with an auxiliary set of representations.

Let us also comment on another contentious issue. Early connectionist networks have been criticized for lacking representational content. This omission made them an easy target for representationalists, who pointed out the connectionist failure to capture general rules. While that criticism has been largely justified, recent developments within the field and advances in deep network architecture have shifted the emphasis of this debate. As contemporary networks work with a limited degree of abstraction, it is unfair to deny them some kind of representational structure. Yet the essential problem remains: as the examples of adversarial attacks illustrate, the representational structures of networks are very distinct from representations in the human mind. When shaped up by content from experience, they become dependent on non-robust features.

4.1 Failures of empiricism

Quite a bit has changed in the outline and efficacy of neural networks over the last few decades. We are now facing a much more aspiring field, and the successes of deep networks across many areas have not gone unnoticed. As a consequence, there is renewed scholarly interest in capturing their philosophical significance, especially in their justification of empiricism. We have seen that an inclusion of convolution led attempts to resuscitate networks' support for the theory. Buckner (2018) argues that, by an automated creation of categories solely from inputs, networks perform transformational abstraction, fully in line with the spirit of empiricism. Convolutional networks apparently support the original empiricist conviction of the real possibility to derive cognitive categories exclusively by extensive processing of a sufficient amount of empirical data. Yet, as evidenced by the adversarial attacks, as well as problems with abstraction and non-transferability of algorithms, processes within networks do not resemble those we encounter in the human mind. If a network aims only at optimizing performance on a given task, its creation of an unusual representational structure presents no major problem. The difficulties arise when the parable of optimizing features is stretched outside of its intended target domain.

Drawing conclusions about human cognitive structures and processes from networks presents an example of such a stretch. Debates on empiricism started off as queries about the constitution of the human mind. The fact that networks generalize in their own unique way provides little support for the resolution of disputes about the workings of the mind or (general) intelligence. Given how distinct their resulting functions are from human methods of abstraction, knowledge transfer and causal reasoning, it is impossible to use them as means of justifying philosophical empiricism. It is appropriate for the advocates of model-free deep networks to argue that their nets offer solutions to a particular problem, within a given data set, but not much beyond that.

Interestingly, even their defenders of the networks are aware that to match the human mind, additional structural features need to be fitted into the system. They speak of an introduction of various priors that influence the net's functionality. Their notion of priors is importantly distinct from the one standardly used in probability theory. Priors encompass any factors that are fixed within the network and precede its actual operations. As will be noted below, this usage comes closer to that of I. Kant and his system of categories. Ilyas et al. (2019) offer "enforcing a prior over the features learned by the classifier" (p.10) as the method for approximating artificial networks to the human mind. Even Buckner (following Goodfellow et al. 2016) admits that to achieve a robust system of categorization, networks would have to adopt "infinitely-strong, domain-general priors".[19] The broadly-conceived priors might come in the form of modules and settings that amend existing model-free networks to match corresponding human capabilities more closely. Yet the addition of such priors is a far cry from empiricism. This admission is especially surprising for Buckner as it reverses his original commitments to empiricism. Once he allows "infinitely-strong" priors to play a prominent role within the system, it is hard to understand why such a system would deserve to be called empiricist in the first place.

Buckner is not the only one voicing concerns about prospects of purely empiricist approaches to categorizations. In fact, there is a line of criticism that denies the very starting point of the argument in the support of empiricism by deep network results. In his critical analysis, Marcus (2018b) has been very vocal in pointing out various obstacles to call networks model-free and, thereby, empiricist. He demonstrates how various crucial components of successful networks are not an outcome of pure processes of generalization, but instead result from a series of conscious decisions of designers and their endless tinkering with the nets' architecture. From an inclusion of a search algorithm[20] to various decision trees and a chosen number of internal layers,

[19] For example, in the case of categorization, one such prior could be the setting that rules that "contrast levels tend to be similar in nearby pixels" to facilitate detection of "coherent ensembles of feature presentations" (Buckner 2018, p. 5362).

[20] See the debate on the Monte Carlo search within supposedly model-free AlphaGo network (Marcus 2018a, p. 7).

a network is not pristine tabula rasa. Even the choice of a general network setup, its depth and connection density are decisions that come prior to its training and as such have little to do with the empirical content. Every such a choice of parameters goes contrary to the empiricist spirit of model-free architecture. In fact, the whole point of Marcus' paper is to deny that model-free networks deserve their name. There are too many inherent designed features within them to call them model free.

One should not overlook differences between the critical stance of Marcus, who sees the entire enterprise of model-free approaches as misguided in its proclaimed ambitions and a modest admission of Buckner who is aware of the need to adjust parameters within networks. It might be useful to come back to the towering figure of Kant to appreciate these differences.

4.2 Kantian approach

Empiricism has, at least since Kant, largely discarded its vision of the mind as an empty slate filled in only with empirical content and few rules of associations. This early conception, which, as we have seen, still has certain appeal among advocates of model-free approaches, was replaced with the empirical fine-tuning of a system that possesses strong initial setup. It looks like Quine is also relying on this reading of empiricism when he speaks of innate dispositions (Sect. 1.4). Initial parameters ("priors" in broad Marcusian sense) can be read as implicit models that streamline empirical content. For the original Kantian approach of the mind, the number of priors remained very small (he introduces twelve very abstract categories of human cognition, such as unity, necessity, and negation, see Kant 1781, A80/B106). It is worth of noting that Kantian categories have direct bearing on the content of cognition: they specify intrinsic limitations on how one can cognize. To use a metaphorical expression, Kant limits the space of possibilities for our "software". The crux of the current debates on the empiricist foundations have shifted our attention on the "hardware" side as it also appears very important in shaping the cognitive capabilities of a system. As Marcus demonstrates, to match human cognition, networks call for a significant number of different priors. Chosen algorithm and network architectures are already setting limitations on their outcomes (Marcus 2018a). Without setting these parameters correctly, network would not function at all. There are also external adjustments to the network during the training process. Yet Marcus' criticism is more dimensional. He sees virtually all tinkering by the network's designers as moving against the spirit of model-free empiricism. We share his conviction that instead of model-free systems, deep networks present fine-tuned cognitive systems, full of various priors. Baroni illustrates the same point on the case of linguistic processing by networks:

> Modern sequence-processing networks are complex systems, equipped with strong structural priors such as gates, encoding and decoding modules and attention. They should not be thought of as "tabulae rasae", as they often were in early debates on connectionism. At the same time, the "innate" biases they encode are rather different from those assumed to shape human linguistic competence. Some researchers are trying to inject into modern networks priors closer to those traditionally postulated by linguists, such as a preference for hierarchical tree structures. … neural networks might solve complex linguistic tasks, but not in the way we expect them to be solved. (Baroni 2020, p. 3–4).

While empirical experience still shapes cognition and endows it with particular content, the resulting picture bears little resemblance to the mind that empiricism is likely to embrace. We hasten to add that *the more priors one has to incorporate into a system, the less likely that system deserves the empiricist label*.

4.3 Beyond empiricism

With inherent limitations of model-free networks that we have mapped, and empiricism largely out of the picture, there remains a further question about how we can ensure that networks resemble the human mind more closely. It is important to stress that a demise of universalist aspirations of model-free approaches should not be interpreted as a signal to return to purely representationalist systems. There is no desire to go back to the fully representational models in the spirit of expert systems or other fully explicit computational approaches. While they might have historically proved very successful in various domains, their inflexibility and thirst for explicit knowledge representations make them unlikely candidates for general cognitive systems.

After the failures of both fully representational and model-free systems, a middle road stays open in the domain of creating artificial cognizers. We do not want to commit ourselves to a particular approach that might be necessary for such a project. A conclusive answer to this outstanding problem requires substantive empirical research and while some of it has already been conducted, a lot more needs to be done. However, a general consensus seems to be emerging. In the words of Pearl (2018): "human-level AI cannot emerge solely from model-blind learning machines; it requires the symbiotic collaboration of data and models". This view is supported by others in the field. Geffner (2018) defends an integration of model-free and model-based approaches (to use his terminology, "learners and

solvers"). Analogous programmatic proclamations have been also made by Spelke and Kinzler (2007) and Marcus (2018b). Marcus specifically defends the need to include several "computational primitives" that help with various cognitive tasks deep learning handles only with difficulties. In his (2018a) paper he recalls a debate with Yann LeCun where he talked about the issue: "At my October 5, 2017 debate with Yann LeCun, I had an opportunity to draw up a preliminary list. The list I proposed was, roughly, the union of a set of computational primitives that I had advocated for in my book The Algebraic Mind (Marcus 2001), and a set of conceptual primitives drawn from Elizabeth Spelke's work on cognitive development (Spelke 1994): representations of objects, structured, algebraic representations, operations over variables, a type-token distinction, a capacity to represent sets, locations, paths, trajectories, obstacles and enduring individuals, a way of representing the affordances of objects, spatiotemporal contiguity, causality, translational invariance, capacity for cost–benefit analysis". The response of his opponent was highly unorthodox: "Provocatively, LeCun argued (when pushed by the moderator, David Chalmers) that none of these need be innate."

These are not just ideological advocacy calls. On the contrary, they result from a long history of research on the mechanisms and content of human minds these authors conducted. Recently, Lake et al. (2017) argued that the inclusion of human-like modules, such as those for folk psychology or folk physics, is advisable for obtaining network performance proximal to that of humans. They argue that the presence of both approaches is critical for cognition: "The difference between pattern recognition and model building, between prediction and explanation, is central to our view of human intelligence" (Lake et al. 2017, p. 3) and as such it has to be preserved in artificial systems as well. The authors also believe that analogous enrichment of the networks facilitate learning efficiency in artificial systems: "if deep neural networks could adopt … compositional, hierarchical, and causal representations, we expect they could benefit more from learning-to-learn" (ibid., p. 17). All these claims serve as the final nail to the coffin of the empiricist program and very much continue the discussion on how to overcome the poverty of stimulus (Sect. 1.4). Hopefully, we have finally learnt that purely empiricist strategies can neither explain human behavior nor sanction human-level efficiency in machine learning.

Still, precise arguments about the form and scope of various cognitive models and forms of integrations with neural networks are open to further inquiry. They cannot be resolved by armchair theorizing, though some inspirations might be suggestive despite being only shallowly anchored in empirical evidence. One such a framing idea is put forward by Geffner (2018) when he argues for the need to integrate model-free and model-based approaches in machine learning because this integration is widely observed in operations of System 1 and System 2 of human minds (see Kahneman 2013). If this integration of dual systems of processing is successful in us, there is a good reason to presume it will be beneficial to artificial systems as well.

5 Conclusion

We have aimed to demonstrate the incapability of purely empiricist approaches in properly tackling cognitive learning processes. Our aim was to show that, despite several historical waves of arguments to the contrary, empiricism cannot be a reasonable candidate for the answer to the question about the workings of the mind.

We began with a philosophical debate on the nature of learning, originating in the 1950s. The core of Chomsky/Quine debate puts restraints for how far one can go with an empiricist account for learning based solely on associations between input data. Behaviorism of Quine is an excellent example of the late strict empiricism. His views run into difficulties when confronted with empirical facts on language learning and understanding abstract concepts. When Quine is pushed to the corner by Chomsky, he unwillingly admits to the presence of some internal structures on shaping up empiricist input, thereby coming closer to the representational side of the dispute. Still, Quine considers his adjusted theory as a novel version of empiricism, thereby seemingly making just a small concession to the original empiricist credentials. For us his concession is just a first step in the eventual demise of empiricism.

In the 1980s, the alleged support for empiricism has been offered by computer scientists within the new field of distributive processing. Early connectionist networks, while built upon then novel ideas, just did not deliver on their promise to bring forward a data-driven artificial counterpart of human intelligence. In their defense, one might argue that back then, computing power was insufficient, and methods of networks' design have not been developed enough. In any case, defenders of early networks were not very successful at fending off computationalist attacks on issues such as compositionality and abstraction. Early networks could not generalize outside of their limited domain and frequently committed mistakes that were unlike anything we have seen in humans (see, for example, the debate on the English past tense in Bullinaria 1994). Genuine successes only came with an adoption of architectures that deserted the original purely empiricist tendencies by adding task-specific dedicated modules.

Nowadays, similar arguments are put forward by the advocates of newest offspring of the artificial intelligence research, deep learning networks. Their robustness and sophisticated connectivity make them ideal candidates for

genuine empiricist models of the mind. Yet, as we have demonstrated, the latest iterations of neural networks systematically fail in processes of abstraction, knowledge transfer, causal reasoning and proximity of their categorization structure to those of humans. We have shown how all these problems stem from the empiricist principles, upon which the networks are anchored. It appears that the dream of empiricism is coming to its resolute end. The only way to move forward on the issue of the artificial mind is to use hybrid models that merge the best of the two worlds. On the one hand, we definitely need content-driven empirical knowledge of the world that model-free approaches rely on. On the other, that empirical content has to be categorized and worked upon by the precise models that ensure the resulting output is sufficiently similar to our own cognizing. While some such hybrid models were historically judged as empiricist, currently known requirements of complexity and extensiveness of these models make the empiricist label very inappropriate.

We close by noting that it should be of a remarkable interest to philosophers that we are entering an era when philosophical quarrels, such as those between empiricism and representationalism, can be empirically tested. Neural networks constitute powerful data-hungry systems that modify themselves in accordance with their inputs. They thereby offer a prime experimental playfield for the justification or refutation of empiricism. As arguments in this paper indicate, there are good reasons to remain skeptical with regards to networks' unambiguous support for the fundamental empiricist thesis. Despite a wave of enthusiasm among computer scientists to the contrary, we find prospects of neural networks that operate only on bare data more of a utopian dream than reality. This is not to say that the entire agenda of machine learning has no bearing on the issue of the foundations of the mind. In fact, it forces philosophers to refine their positions and rethink the precise content of their theses. Yet as long as influential computer scientists keep making strong and unjustified claims about powers of their systems, we can only recommend them to get better acquainted with the history of philosophical thinking and skeptical voices within their own community to avoid some of the pitfalls they tend to place themselves into.

Acknowledgements We want to thank the anonymous reviewers for providing a detailed commentary on earlier drafts of the paper. The work on this article was supported by the Grant GA16-15621S of the Czech Science Foundation.

References

Aizawa K (2003) The systematicity arguments. Kluwer, Dordrecht
Ashok P et al (2020) DeepAbstract: neural network abstraction for accelerating verification. arXiv:2006.13735
Baroni M (2020) Linguistic generalization and compositionality in modern artificial neural networks. Philos Trans R Soc B 375:20190307
Bromley J, Bentz JW, Bottou L, Guyon I, LeCun Y, Moore C et al (1993) Signature verification using a "siamese" time delay neural network. Int J Pattern Recognit Artif Intell 7(04):669–688
Buckner C (2018) Empiricism without magic: transformational abstraction in deep convolutional neural networks. Synthese 195:5339–5372
Bullinaria J (1994) Learning the past tense of English verbs: connectionism fights back. Edinburgh University Technical Report – May 1994, available at https://www.cs.bham.ac.uk/~jxb/PUBS/PTEV.pdf
Burrell J (2016) How the machine 'thinks': understanding opacity in machine learning algorithms. Big Data Soc 3(1):1–12
Carnap R (1928/1967) Der logische Aufbau der Welt. Trans. by Rolf A George as The Logical Structure of the World. Berkeley: University of California Press
Cho K, Van Merrienboer B, Gulcehre C, Bougares F, Schwenk H, Bengio Y (2014) Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)
Chomsky N (1967) Review of B F Skinner's *Verbal Behavior*. In: Jakobovits LA, Miron MS (eds) Readings in the psychology of language. Prentice-Hall, Hoboken, pp 142–143
Chomsky N (1968) Quine's empirical assumptions. Synthese 19(1/2):53–68
Chomsky N (1979) Language and Responsibility Pantheon. Trans. John Viertel (Based on conversations with Mitsou Ronat.) Reprinted in Chomsky N (2007) *On Language* New Press
Cowie F (2017) Innateness and Language. In: Edward NZ (ed) The stanford encyclopedia of philosophy (Fall 2017 Edition). https://plato.stanford.edu/archives/fall2017/entries/innateness-language/. Accessed 10 Jan 2020
Darwiche A (2018) Human-level intelligence or animal-like abilities? Commun ACM 61(10):56–67
Eslami SM, Heess N, Weber T, Tassa Y, Kavukcuoglu K, Hinton GE (2016) Attend, infer, repeat: fast scene understanding with generative models. Presented at the 2016 Neural Information Processing Systems conference, Barcelona, Spain, December 5–10, 2016. In: Lee DD, Sugiyama M, Luxburg UV, Guyon RI (eds) Advances in Neural Information Processing Systems 29 (NIPS 2016). Garnett, pp. 3225–33. Neural Information Processing Systems Foundation
Fodor J (1992) Theory of content and other essays. MIT Press
Fodor J (1998) Concepts. Oxford University Press
Fodor J, Pylyshyn Z (1988) Connectionism and cognitive architecture: a critical analysis. Cognition 28:3–71
Garnelo M, Shanahan M (2019) Reconciling deep learning with symbolic artificial intelligence: representing objects and relations. Curr Opin Behav Sci 29:17–23
Geffner H (2018) Model-free, model-based, and general intelligence. arXiv preprint, arXiv:1806.02308

Gershman SJ (2019) The generative adversarial brain. Front Artif Intell 2(18):1–8. https://doi.org/10.3389/frai.2019.00018
Gibson RF (1987) Quine on naturalism and epistemology. Erkenntnis 27(1):57–78
Huang S, Papernot N, Goodfellow I, Duan Y, Abbeel P (2017) Adversarial attacks on neural network policies. arXiv preprint arXiv:1702.02284
Ilyas A, Santurkar S, Tsipras D, Engstrom L, Tran B, Madry A (2019) Adversarial examples are not bugs, they are features. arXiv preprint arXiv:1905.02175
Johnson K (2004) On the systematicity of language and thought. J Philos 101(3):111–139
Kahneman D (2013) Thinking fast and slow. Farrar, Straus and Giroux, New York
Kant I (1781/1958) *Critique of Pure Reason*, Norman Kemp Smith (trans.), London: Macmillan
Kemp G (2006) Quine: a guide for the perplexed, Continuum
Kulkarni TD, Narasimhan KR, Saeedi A, Tenenbaum JB (2016) Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation. arXiv preprint http://arxiv.org/abs.1604.06057
Lake BM, Salakhutdinov R, Tenenbaum JB (2015) Human-level concept learning through probabilistic program induction. Science 350(6266):1332–1338
Lake BM, Ullman TD, Tenenbaum JB, Gershman SJ (2017) Building machines that learn and think like people. Brain Behav Sci. https://doi.org/10.1017/S0140525X16001837
Lake B, Baroni M (2018) Still not systematic after all these years: on the compositional skills of sequence-to-sequence recurrent networks, ICLR conference page. Accessible at https://openreview.net/forum?id=H18WqugAb
Marcus GF (2001) The algebraic mind. MIT Press, Cambridge
Marcus G (2018a) Innateness, AlphaZero, and artificial intelligence. arXiv preprint arXiv:1801.05667
Marcus G (2018b) Deep learning: a critical appraisal. arXiv preprint arXiv:1801.00631
Mayor J, Gomez P, Chang F, Lupyan G (2014) Connectionism coming of age: legacy and future challenges. Front Psychol 5:187
Ng A (2016) What artificial intelligence can and can't do right now. Harvard Business Review, 9(11)
Norvig P (2017) On Chomsky and two cultures of statistical learning https://norvig.com/chomsky.html. Accessed 4 May 2021
O'Brien G, Opie J (2009) The role of representation in computation. Cogn Process 10(1):53–62
Och FJ (2002) Statistical machine translation: from single-word models to alignment templates. Dissertation, RWTH Aachen
Pearl J (2018) Theoretical impediments to machine learning with seven sparks from the causal revolution. arXiv preprint arXiv:1801.04016
Pylyshyn Z, Demopoulos W (eds) (1986) Meaning and cognitive structure. Ablex Publishing, New Jersey
Quine WV (1950) Identity, ostension, and hypostasis. J Philos 47(22):621–633
Quine WV (1969a) Natural kinds. In: Rescher N (ed) Essays in Honor of Carl G Hempel. Synthese Library (monographs on epistemology, logic, methodology, philosophy of science, sociology of science and of knowledge, and on the mathematical methods of social and behavioral sciences), vol 24. Springer, Dordrecht
Quine WV (1969b) Reply to Chomsky. In: Davidson, Hintikka (eds) Words and objections. Essays on the work of W V Quine. Reidel, Dordrecht, pp 302–311
Quine WV (1972) Methodological reflections on current linguistic theory. In: Davidson D, Harman G (eds) Semantics of natural language. D. Reidel Publishing Co., Dordrecht, pp 442–454
Quine WV (1975) The nature of natural knowledge. In: Guttenplan SD (ed) Mind and language. Clarendon Press, pp 67–81
Quine WV (1976) Linguistics and philosophy. In: Quine WVO (ed) The Ways of Paradox, and other essays, revised and enlarged ed. 1976. Harvard University Press, Cambridge, pp 56–58
Quine WV (1978) Facts of the Matter. Southwestern J Philos 9(2):155–169
Ramsey W, Stich S (1991) Connectionism and three levels of Nativism. In: Fetzer JH (ed) Epistemology and cognition. Studies in cognitive systems, vol 6. Springer, Dordrecht
Rougier NP (2009) Implicit and explicit representations. Neural Netw 22(2):155–160
Rumelhart DE (1998) The architecture of mind: a connectionist approach. In: Thagard P (ed) Mind readings: introductory selections on cognitive science. MIT Press, pp 207–238
Rumelhart DE, McClelland JL (1987) Parallel distributing processing, vol 2. MIT, Cambridge
Russell B (1914) Our knowledge of the external world. Open Court Publishing, La Salle (Reprinted Routledge, London and New York, 2000)
Schmidhuber J (2015) Deep learning in neural networks: an overview. Neural Netw 61:85–117
Silver D, Schrittwieser J, Simonyan K, Antonoglou I, Huang A, Guez A et al (2017) Mastering the game of Go without human knowledge. Nature 550(7676):354–359
Skinner BF (1957) Verbal behaviour. AppletonCentury-Crofts, New York
Spelke ES, Kinzler KD (2007) Core knowledge. Dev Sci 10(1):89–96. https://doi.org/10.1111/j.1467-7687.2007.00569.x
Sun S et al (2016) On the depth of deep neural networks: a theoretical view. In: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16), pp 2066–2072
Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, Fergus R (2013) Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199
Tan C, Sun F, Kong T, Zhang W, Yang C, Liu C (2018) October) A survey on deep transfer learning. International conference on artificial neural networks. Springer, Cham, pp 270–279
Taylor P, Hobbs J, Burroni J, Siegelmann HT (2015) The global landscape of cognition: hierarchical aggregation as an organizational principle of human cortical networks and functions. Sci Rep 5(1):1–18. https://doi.org/10.1038/srep18112
Van Cleve J (1999) Problems from Kant. OUP, New York
Walker SF (1992) A brief history of connectionism and its psychological implications. In Clark, A and Lutz, R (eds) Connectionism in Context, Springer-Verlag, Berlin, pp 123–144
Zednik C (2019) Solving the black box problem: a normative framework for explainable artificial intelligence. Philos Technol. https://doi.org/10.1007/s13347-019-00382-7
Zhou Z, Firestone C (2019) Humans can decipher adversarial images. Nat Commun 10(1):1–9