Minds and Machines (2020) 30:47–75
https://doi.org/10.1007/s11023-020-09519-6

ORIGINAL PAPER

A Puzzle concerning Compositionality in Machines

Ryan M. Nefdt[1]

Received: 25 October 2019 / Accepted: 12 February 2020 / Published online: 23 February 2020

Abstract
This paper attempts to describe and address a specific puzzle related to compositionality in artificial networks such as Deep Neural Networks and machine learning in general. The puzzle identified here touches on a larger debate in Artificial Intelligence related to epistemic opacity but specifically focuses on computational applications of human level linguistic abilities or properties and a special difficulty with relation to these. Thus, the resulting issue is both general and unique. A partial solution is suggested.

Keywords Compositionality · Deep Neural Networks · Deep learning · machine learning · Epistemic opacity · Artificial Intelligence

1 Introduction

The principle of compositionality is a widely endorsed claim about the nature of natural language semantics. Although its precise definition is a matter of debate, its intuitive appeal is unassailable within linguistics, cognitive science, and philosophy. The present work aims to question its further applicability to the field of Artificial Intelligence (AI) with specific focus on research into machine learning and artificial neural network architectures. In this paper, I present a puzzle for applying the principle of compositionality to artificial networks such as Deep Neural Networks (DNNs) with hidden layers. The puzzle concerns the identification of meaningful parts, which I argue is essential for semantic compositionality, within the overall network structure.

There has been a recent upsurge in work related to whether or not artificial networks used in machine learning can indeed capture compositionality (often assumed to be a property of human language competence) (Pinker 1984; Fodor and Pylyshyn 1988; van Gelder 1994; Marcus 2003; Baroni 2019). However, it is not always clear what the exact target of these computational models are (compositional output or

✉ Ryan M. Nefdt
Ryan.Nefdt@uct.ac.za

[1] Department of Philosophy, University of the Cape Town, Cape Town, South Africa

compositional input or something in the process itself). Therefore the layout of the paper is as follows. In Sects. 2 and 2.3 respectively, I describe the principle of compositionality discussed in the literature and aim to clarify some conceptual confusion surrounding the concept. In Sect. 3, I discuss the problem of epistemic opacity in AI more generally before presenting a unique puzzle related to Artificial Neural Networks and their application to natural language in Sect. 4. I conclude with a suggestion of where one might begin with an answer to this puzzle.

2 Compositionality

As mentioned at the onset, there is no universally agreed upon definition of compositionality for natural language. Although the statement thereof, the so-called principle of compositionality, is often referenced, its precise interpretation varies among theorists. In this section, I will not aim to provide the much-sought after definition, rather I will attempt to identify one necessary component of the principle as it is generally understood in linguistics and philosophy. In other words, whichever interpretation you prefer needs to possess this component on pain of omitting something essential about compositionality, namely the concept of a ‘meaningful part’.

Essentially, compositionality concerns the nature of semantic derivation. It states that the meaning of a whole expression is determined by or somehow dependent on the meaning of its parts and their combination. This is taken to mean that any change in the whole is the result of some change in the parts or their combination exclusively. Many semanticists, following the Montagovian tradition, interpret this claim functionally, i.e. the meaning of the whole is a function of the meaning of its parts and their combination. We’ll follow some of the lines of this narrative here.

2.1 Some Background

The genesis of the principle of compositionality (PoC) has been linked with the writings of Frege, hence the term “Frege’s principle” (sometimes used synonymously). In ‘Sinn und Bedeutung’ (1908), Frege challenged a simple notion of compositional meaning in terms of co-reference (due to Mill), by testing the latter view on the so-called substitution thesis [often erroneously assumed to be equivalent to PoC, see Szabó (2000)]. Yet his distinction between sense (*Sinn*) and reference (*Bedeutung*) did aim to rescue a compositional account of meaning in some other form, as the compositionality of sense.[1]

[1] Janssen (2012) argues that Frege was not the source (nor an adherent) of the PoC. In fact, he argues that Frege subscribed to a quite different principle for natural language semantics. Its true origins can actually be traced further back than Frege to Lotze, Wundt and Trendelenburg, according to Janssen. Hodges (2012) goes further to trace the concept to the works of the tenth century Arab scholar Al-F$\bar{a}$r$\bar{a}$b$\bar{i}$ who could have in turn found it in 3rd century commentaries on Aristotle.

The modern idea of the principle can be found in Montague (1974) and Partee (2004) among others and it usually takes the form of the following type of statement, let's call it the intuitive principle of compositionality:

> (iPoC): The meaning of a complex expression is determined by the meaning of its component parts and the way in which they are combined.

In the methodology of logic and computer science, it has been considered the standard way of interpreting formal or programming languages (although alternatives do exist).[2] Tarski's (1933) definition of truth for formal languages has a natural compositional interpretation (Janssen 2012). Davidson (1967) used what he called Tarski's T-schema as a basis for a compositional semantics for natural language. In the twentieth century the principle was widely adopted in the philosophy of language and logic, through Carnap, Quine, Davidson and various others. Now, it has become an essential part of most linguistic theories including generative approaches such the seminal Heim and Kratzer (1998).

Unfortunately, there is still no consensus on the correct definition of compositionality in natural language. Furthermore, it is not clear if it is exclusively a methodological principle (Dever 1999) or can be empirically tested (Dowty 2007).

Consider the iPoC again. The statement as it stands is both vague and in need of clarification. The problem is that there does not seem to be a neutral way of going about this clarification.

Nevertheless, in this section, I will attempt to stay as neutral as possible. Starting with the term "complex expression" which will be characterised as a syntactic object built up from simple (and perhaps other complex) expressions, as will its subcomponents. We will remain characteristically reticent about meanings and what is exactly meant by 'meaning' for the moment.

"Determined by" is usually interpreted functionally, which suggests that given a syntactic object as an input, a semantic object or value is returned as an output. One might worry that this suggests a unique semantic output for every syntactic expression but in natural language this tends to overgenerate as there are distinct expressions which arguably should be assigned the same meanings. That would be is a very strong constraint on meaning. However, such a scenario is not an issue for compositionality. Consider the sentences below:

1. Jimmy threw the ball.
2. The ball was thrown by Jimmy.

[2] Propositional logic is a good example of a formal language with a simple compositional semantics. The meaning of a formula is a truth value and the meaning of a complex formula is a function of the meanings/truth values of its components. Predicate logic is not as simple a matter. Following Pratt (1979), we know that "there is no function such that meaning of $\forall x\phi$ can be specified with a constraint of the form $\mathcal{M}(\forall x\phi) = F(\mathcal{M}(\phi))$" (Janssen 1997: 498). In other words, the meaning of a universally quantified formula is not straightforwardly given in terms of a function from the meaning of its parts, at least not by means of the standard Tarskian interpretation.

The sentences above both seem to express the same meaning but consist of different lexical items such as the preposition *by* as well as a different method of combination (active vs passive). It is quite apparent from the literature that if the term 'function of' (when used to interpret 'determined by') is to be conceived of in its strict mathematical sense, it involves a surjective function.

A stronger interpretation has it that we start from atomic elements and assign a meaning to each of those, then define a semantic rule for every syntactic rule. In this way, we have a compositional semantic procedure which parallels every syntactic one (this is in essence Montague's homomorphism definition), also called "parallelism". This approach does not overgenerate, since the rules that combine (1) and (2) may be different but along with the meaning of the words they could produce the same meaning for the expressions. The functions sqrt and cbrt will both produce the number 2 when the input is 4 or 8 respectively. We will be more precise about these definitions in Sect. 2.3.[3]

Two worries might arise here, one syntactic and the other semantic. From the syntactic perspective, one might be concerned that a ban on identical syntactic expressions with nonidentical meaning follows from the simple notion of compositionality. "If a language is compositional, it cannot contain a pair of non-synonymous complex expressions with identical structure and pairwise synonymous constituents" (Szabó 2007). This amounts to a ban on ambiguity at the sentence level. Pseudo-conditionals appear to be in conflict with this condition. Consider the (3) and (4) below from Kay and Michaelis (2011):

3. If you're George W. Bush, you're now allowed to lie in the faces of trusting young voters.
4. If you're pleased with outcome, you may feel like celebrating.

It is argued that (3) and (4) seem to have the same syntactic structure and yet (3) does not express a hypothetical statement of any kind while (4) is a conventional conditional which does. The semantics of these constructions seems to be quite different. "[N]o hypothetical situation is posed; it appears that a categorical judgement is expressed [...] and the subject of that judgement is not the addressee but the person identified as x [George Bush]" (Kay and Michaelis 2011: 2). However, it is not clear from their example that a hypothetical reading is not possible for (3). The mere fact that (3) is truth-conditionally equivalent to a categorical statement is not a good argument for such a claim. In fact, so-called "biscuit conditionals", in which a conditional reading is much less available, might make the point more clearly, as in (5) below:

5. There are biscuits on the sideboard, if you want some.

[3] Parallelism has other weaknesses though. For instance, it strongly suggests a building metaphor of a step-by-step procedure mapping syntactic combination with semantic interpretation. Possible world semantics does not respect this constraint, nor do semantic formalisms with intermediary representations like Montague's Type 2.

The more semantic worry is brought out by the much-discussed cases of scope ambiguity. Consider the example below:

6. Every boy loves some girl.

There are at least two possible readings of this sentence. The first is that every boy loves some girl in the sense that each boy loves a distinct girl. The second is that there is one girl who is extremely popular. Given our definition of compositionality, we seem to have a violation since the components and method of combination are the same and yet the resulting semantic analyses differ.

In terms of the former problem, this situation can be resolved by interpreting the statement about method of combination as involving the combination of the meanings of the components and not the syntactic components themselves.

> [This] permits the existence of non-synonymous complex expressions with identical syntactic structure and pairwise synonymous constituents, as long as we have different semantic rules associated with the same syntactic rules (Szabó 2012: 70).

The problem of scope ambiguity has been amply addressed in many ways, ranging from alternative syntactic combination (the Montagovian solution) and type shifting to Cooper Storage (see Cooper 1975) or underspecification. Basically, both strategies allow for the components to be combined differently (albeit at different levels) and thus map onto different semantic rules of composition (or at least to be applied in a different order). Some question the compositionality of the latter solutions (see Lappin and Zadrozny 2000) but I shall obviate that discussion here.

Lastly, what is meant by "component" as it is used here? The convention in the literature is to take this to mean *constituent*. In linguistics, this is a loaded term. Usually, it refers to sets of linguistic items which act as structural units in given expressions. There are various tests for constituency such as coordination, deletion, modification etc. In terms of representation, constituents are the groupings of items that appear in the hierarchical tree diagrams of phrase structure grammar. Consider the sentence (7) below:

7. The host speaks to the caterer after the ceremony.

This sentence can be separated into distinct constituents. One can follow phrase structure syntax, in which roughly the *NP*—the host, the *VP*—the host speaks, *PP*—to the caterer would be constituents. Alternatively, the verb phrase or predicate *speaks to* and its two arguments *the host* (subject) and *the caterer* (complement) could be considered constituents. This latter strategy is common to formalisms such as dependency grammar (see Rambow and Joshi 1992). Items such as *After the ceremony* are sometimes called adjuncts and can float independently of the other constituents (as opposed to complements which are obligatory).

Linguists often interpret compositionality and "components" as not only involving constituents but immediate constituents or constituents immediately governed by the node above (i.e. daughters not granddaughters). Immediate constituents "appear at the first level in the analysis of the form into ultimate constituents" (Hodges 2012: 249) or the terminal alphabet at the bottom of the syntactic tree. Therefore, our tentative definition can be modified to incorporate the clarifications mentioned in this section, let's call it the functional PoC.

> (*f*PoC): The meaning of a complex syntactic expression is a function of the meanings of its immediate constituents and the syntactic rule used to combine these constituents.[4]

There is of course much more to say on this topic. How is this relationship to be mathematically represented (often read as 'what kind of formal mappings capture the PoC')? Do we need immediate constituents (Szabó 2012)? Does natural language even exhibit compositional structure or do humans learn complex partially saturatable constructions (see Croft 2001; Goldberg 2015; Jackendoff 2002)? How does compositionality relate to the productivity and systematicity of natural language? The list goes on. But for our present purposes, this definition highlights one core component of the PoC more generally, the notion of parthood *via* constituency.

2.2 'Meaningful Parts'

Whatever the interpretation of the PoC, or terms such as "determined by" [see Szabó (2000) for a supervenience reading thereof] or constituent (which differs within syntactic formalisms such as dependency grammar) or even syntactic combination/rule, one notable presupposition of the principle is that complex expressions exhibit a part-whole structure. Werning (2012) calls this "semantic constituency" or "a correspondence relation between the part-whole relation in the linguistic domain and some part-whole relation on the level of meanings" (634).

The insight dates back to Frege, for whom the sense of an expression reveals a mereological structure. "We can regard a Sense as a mapping of a thought: corresponding to the part-whole relation of a thought and its parts we have, we have by and large, the same relation for a sentence and its parts" (Frege 1919: PW 255). This is not the case for reference for Frege.

In order for the meaning of the whole to be generated, determined or rendered, there has to be some notion of a meaningful part. Meaningful parts are like the atoms which combine to form molecules and ultimately chemical compounds. They are identifiable, independent, and separable entities. You could think of these atoms as words in the case of language but there is even considerable compositional structure below the word level. For instance, in derivational morphology, morphemes play the role of "meaningful parts". Consider the word *decomposition* which is composed of three separate morphemes *de*, *compose* and *tion*. Of course, in morphology

[4] I neglected to give an interpretation of what is meant by "syntactic rule" here. This is a matter of theoretical perspective to a large extent. Traditionally, categorial grammars have been used as well as phrase structure grammars. However, the options are without obvious limit.

the meaningful parts are not independent as in the case with words.[5] So what is it for something to be a meaningful part? Outside of Frege's idiosyncratic view on meaning, we would need both a concept of "meaningful" and of "part". Let's start with the latter.

One way to define a part in mereology is in terms of certain formal properties. According to Lesniewski (1916), parthood is generally taken to have three such properties:

a. Irreflexivity: $\neg Pxx$. Nothing is a part of itself.
b. Asymmetry: $Pxy \rightarrow \neg Pyx$. If x is a part of y, then y is not a part of x.
c. Transitivity: $(Pxy \wedge Pyz) \rightarrow Pxy$. If x is a part of y and y is a part of z then x is a part of z.[6]

Of course, this definition would not give us much traction on part *qua* constituent in linguistics. According to the rationale any part of an expression would be a constituent of that expression. This definition quickly runs into triviality. Consider the sentence below:

8. A man walked into the room with nothing good on his mind.

The way in which such a sentence would generally be compartmentalised into constituents in a neutral (as possible) theory of syntax is something like:

$[_x[_y[\text{a man}]\ \text{walked}]_z[\ \text{into}_u[\text{the room}]]]$

Ignoring hierarchical structure and relationships between these syntactic objects for the moment, we can see that the three properties of parthood are respected in this case. This might potentially lend credence to the idea that constituents are just linguistic parts. But parthood is more general than constituency as shown in the perfectly compatible part-whole structure below,

$[_x[\text{a}]_y[\text{man walked}]_z[\textit{into}]]_u[\textit{the}_v[\text{room}]]$

This division is equally compatible with a notion of parthood but it does not dovetail with the idea of constituency. Thus, parthood by itself does not narrow down the class of relevant objects enough. However, here the concept of "meaningful" might help.

At first blush, we might consider syntax to exhaustively demarcate the domain of meaningful parts. Constituents seem to be "natural" groupings of linguistic material

[5] However, the precise definition of word-hood assumed from isolating languages such as English and Chinese is not generalisable to agglutinating languages such as Turkish, Yupik and Nguni languages, partly due to the vague lines between morphology and syntax in these latter families. See Nefdt (2019) for a philosophical view on the difficulty of defining words and Haspelmath (2011) for a linguistic discussion.

[6] In some literature, parthood is defined as a partial ordering, i.e. reflexive, antisymmetric and transitive. This allows a part to be a part of itself which when viewed from the point of view of set theory seems to invite inconsistencies.

which act as units during syntactic processes. For instance, they can allow for movement and deletion,

8a. Into the room, a man walked with nothing good on his mind.
8b. A man walked ~~into the room~~ with nothing good on his mind

Whereas neither **Man a walked room into the with nothing good on his mind* nor **A man walked ~~into~~ the room with nothing good on ~~his~~ mind works*. However, not all syntactically relevant groupings are semantically relevant. To see how this is the case, consider the *Extended Projection Principle* (Chomsky 1982) which states that languages such as English must possess a subject or more specifically that subjects are mandatory in *DP*s (determiner phrases) even when there is no semantic subject or agent in the surface form. Sentences like *It is raining*, in which "it" is semantically vacuous, are good examples (the phenomenon of "do-support" in English is another example).

Dever (2012) describes the situation with the PoC as a "screening off" process from the lexical to the sentential level.

> General considerations of the supervenience of the features of wholes on the features of their parts do not suffice here: we are asking for determination not by all properties of the parts, but only by the specifically semantic properties of the parts; and we are asking for determination of the meanings of the complex expressions, and meanings are extrinsic features of expressions, and extrinsic features are typically not determined by features of parts (92).

We do not have to answer the loaded question of 'what is meaning?' to make sense of the claim here. You could have a separate account of meaning as use (Wittgenstein 1953), or inferential role (Brandom 1994),[7] concepts (Jackendoff 2002) or internal mental instructions (Pietroski 2018). The essential element to all of these things is that compositionality requires some concept of a meaningful part, whatever meaning turns out to be.[8]

Thus, I suggest that *meaningful parts* are going to be those constituents which play a significant role in semantic composition. "It" in the sentence "It is raining" plays no significant semantic role. The *role* the part plays determines its meaningful status, not only individual features of the parts or syntactic constituents themselves. Here the notion of *meaning* is reduced to an instrumental value. In other words, the question becomes which parts are useful to the calculation of the meaning of whole expression? Note, these parts could be syntactic constituents, morphemes or

[7] Inferentialism's "top-down" notion of compositionality might not naturally dovetail with some of the remarks made here. They tend to take the sentence as the primary unit of meaning and derive subsentential semantic value from there. Specifically, Brandom's account sees language as recursively structured but doesn't see meaning as compositional. See Brandom (2007) for more. I thank Bernard Weiss for this observation.

[8] Consider Jabberwocky sentences or Chomsky's *Colorless Green Ideas Sleep Furiously*, even in the absence of knowing what the meaning is, we can still identify what the meaningful parts are (or should be).

partially productive constructions (such as those found in some idioms). Their membership as semantic or meaningful parts, however, is not determined exclusively by these latter features. In order to be classified as a meaningful part, an item needs to respect the properties of parthood as well as play a significant role in the meaning of the whole expression (whatever you take "meaning" itself to be).

The view presented here does not automatically contradict strong versions of the PoC such as parallelism or Montague's rule-to-rule mappings since semanticists can (and do) take these mappings to be represented in terms of partial morphisms of some sort. This modification would allow for certain elements of the syntax to receive no semantic value. Further discussion of the precise formal relationship between syntax and semantics is beyond the present scope [see Pagin and Westerstahl (2010) for a few options]. What is relevant here are two principles I take to be relatively uncontroversially related by the PoC, the first is required by all versions and the second is a specific requirement of what I will call "process compositionality" in the next section.

> *Meaningful Parts Principle* (MPP): For the meaning of the whole to be determined by the meaning of its parts (and their syntactic combination), there needs to be meaningful parts.
>
> *The Knowable Parts Principle* (KPP): In order to know the meaning of a whole expression, we must be able to identify what the meaningful parts are.

Note that KPP does not state that we have to know the *meaning* of the parts [which would be the "argument from understanding" discussed in Szabó (2000)], merely that we have to be able to identify what the meaningful parts are by means of the roles they play (or some other mechanism).

2.3 Processes Versus States Versus Outcomes

An important but often neglected distinction in the compositionality debate relates to the different types of compositionality a system can exhibit.

The most common kind of compositionality discussed in the literature (and the discussion above) is what I will call *Process Compositionality*. The central idea is that the property of compositionality is located at the procedural level. What this means is that if a compositional procedure, such as a rule-to-rule mapping, is followed then the system in question is process compositional.

The PoC discussed above assumes that meaningful parts are composed systematically or functionally such that they generate complex or composite meaningful expressions. It specifies a procedure for compositional structures. There are various ways in which process compositionality can be achieved. Jacobson (2002) for instance proposes what she calls "direct compositionality". In this framework, every syntactic constituent must receive a semantic value, or "that for every syntactic operation there must be a corresponding semantic operation" (Barker and Jacobson 2007: 2). This means that even words like "it" in mandatory subject positions in English have semantic values. Unlike some versions of compositionality, which allow for operations to be “held off” until later interpretation in an LF (Logical Form) level, direct compositionality insists on immediate semantic resolution of any syntactic unit. Importantly, direct compositionality is process compositional in the way I am discussing. “[D]irect compositionality is a type of compositionality, where (roughly) a theory of grammar is compositional if the meaning of an expression can be reliably computed from the meanings of its parts” (Barker and Jacobson 2007: 2). In other words, there is a procedure for a system to follow in order to qualify as directly compositional. If the process does not meet this condition, it fails to be such, despite the outcome of the process. For instance, theories which posit quantifier raising (where a quantifier is raised to a higher position in a tree from which it is then interpreted) are not directly compositional.

Another attempt at procedural or process compositionality is found Baggio et al. (2012). There, the authors attempt to investigate the processing consequences of the PoC. They posit that compositionality could be considered as a processing principle given a certain concept of modularity in which the language module is “informationally encapsulated” (in terms of lexicon and syntax) or cognitively impenetrable in the parlance of Fodor (1983) and Pylyshyn (1984). They go on to define incremental composition, which starts “from the observation that ‘function’ in the definition of compositionality needs to refer to some computable input-to-output mapping, and that inputs—lexical meanings and syntactic rules or constraints—must be given incrementally” (Baggio et al. 2012: 659). Most versions of the PoC assume some sort of process compositionality or specification of how the meanings of the parts are composed to generate the whole. We’ll see below that this is also the most difficult type of compositionality to detect in artificial environments as it requires both MPP and KPP.

State compositionality, on the other hand, is a property of a structure identified by the possibility of decomposing that structure or state into smaller meaningful units. A helpful analogy is a puzzle here. A puzzler might have used particular heuristics to construct the overall picture (corners first, left to right, colour matching etc.) yet the state of the completed puzzle can be deconstructed (for later reconstruction, perhaps) in terms of other meaningful arrangements (ignoring the case of randomly deconstructing here). The state of a system itself can be said to compositional in this sense if it can be subdivided into meaningful parts.

However, state compositionality is *theoretically* independent of process compositionality (although in many cases they do coincide). In other words, it should be possible for a compositional process to be followed which results in a non-state compositional state. The obvious cases involve situations in which the process is interrupted or defective in some way. For instance, you could attempt to build an expression from meaningful parts such that at each stage of the process you have composed a meaning (incrementally) but at the end of the process the meaning of the entire expression is not computable from the meanings of those specific parts. Consider the case of a second language speaker who through literal translation stumbles upon an idiomatic expression. In such a case, the conventional meaning can *block* the compositional one. For example, if the speaker were to trying to explain their activity of solving a math problem in a classroom, they might say something like:

9. I solved for *x*, then I went back to the drawing board to solve for *y*.

The situation elicited by (9) is one in which the sentence means, to an English speaker, that the individual started over to solve for *y* (given the conventional meaning of "back to the drawing board"). But of course, the process compositional meaning merely involves going back to a literal drawing board on which the math problem was stated. Although the process might have been compositional, the resulting sentence is interpretable otherwise.

However, these examples might seem ad hoc or exceptional.[9] Evans (1981) distinguishes between two possible internal systems for constructing a 100 sentence list which highlights the more general distinction between process and state compositionality well. The first system contains axioms or primitives while the second contains composition rules and constituents. He argues that the former unlike the latter will be unable to predict the human speaker's ability to understand previously unheard or novel sentences (because it's not a compositional process). The two systems create distinct dispositions, ones which have differing explanatory power. Thus, process compositional structures might create dispositions that allow for the understanding of novel expressions. This might be why machine learning modellers sometimes test for compositional generalisation based on a machine's ability to deal with novel data (as we will see). State compositionality certainly requires MPP but only a Weakened version of KPP in which an identification of "some meaningful parts" not necessarily those used in the actual computation. The axioms can be decomposed but they are used or memorised as whole chunks by the cogniser. In other words, the axiom system is state compositional while the rule-based one is process compositional.

One might still worry that process compositionality necessarily leads to state compositionality. If a process is compositional, it means that the meaning of the whole is derived within a compositional procedure. This would in turn suggest that the compositional procedure was applied to "parts", hence the whole must have been decomposable? How can a compositional procedure (in terms of process compositionality) be applied to a non-decomposable (i.e. a non-state compositional) expression?[10]

This would indeed be a conceptual concern within the framework. However, the separation between process and state compositionality should rather be thought of as one of generation not necessarily application. In this sense, a compositional procedure can generate a non-state compositional state as in the case of the accidental idiom creation. Consider further the case of a chemical reaction precipitated by a catalyst. The resulting covalent bond might be extremely hard to decompose and moreover if decomposition is indeed possible, it would not recover all the original "parts" such as a the catalyst used to initiate the reaction in the first place. The process involved parts which were composed to form

[9] I thank an anonymous reviewer for drawing my attention to this possibility and guiding me to seek out more general examples.

[10] I thank an anonymous reviewer for pressing me on this point.

the whole chemical bond. But the resulting bond is not decomposable into those same parts.

In language, the example of word order freezing witnessed in so-called free word order languages produces a similar pattern. Jacobson (1984) identified the Russian sentence *mat' ljubit doc'* (the mother loves the daughter) as a case which only allows interpretations in which *mat'* is the subject (despite both words having identical nominative and accusative forms respectively). This suggests that the meaning 'the daughter loves the mother' is not decomposable from the whole expression despite involving the same linguistic parts. In other words, the state is not decomposable into a meaning which would be licensed otherwise in terms of the parts. Of course, the process in this case does lead to one possible state which is decomposable ('the mother loves the daughter' reading) but it blocks another ('the daughter loves the mother' reading). For the most part, however, process compositionality does indeed lead to state compositionality.

In the other direction, a structure might be said to be state compositional in the absence of a compositional process. The reason for this is that state compositionality is usually identified by the possibility of decomposition into meaningful parts. Take the case of an arithmetic calculation. A student of mathematics might have used a number of heuristics to perform a particular calculation. The resulting equation, however, could be represented in terms of binary operations on series of 1 s and 0 s. This is due to the fact that compositional structures are *multiply realisable*. In the case of language acquisition, the point might be made more clearly. Consider the following example adapted (for difference purposes) from Szabó (2000).

A child named Arthur learns the sentences 'It is raining' and 'This is an apple' respectively. He is also said to know that the sentence 'Rain is falling' can be used in the exact same circumstances as the former. But he fails to understand the sentence 'This apple is falling' (perhaps through failure to identify picture-sentence pairs or something of the sort). This suggests that Arthur knows the meaning of 'Rain is falling' which is a compositional sentence (not idiomatic or conventional) but fails to apply a compositional procedure for deriving its meaning.

In other words, Arthur processes 'Rain is falling' through meaning association not through individually assigning meanings to each word and functionally composing those meanings. The sentence, for Arthur, is not process compositional. But this does not mean that the sentence is one continuous unit (like a word) for Arthur. He could recognise that it is state compositional, i.e. has parts. To another, more knowledgeable, speaker the sentence would be both state and process compositional.

The only requirement for state compositionality is that the whole expression is decomposable into smaller meaningful parts not that it needed to be built up from those parts. If there is a systematic procedure for decomposing expressions into smaller meaningful parts, this is not to say that they were processed in that way. The related topic of lexical decomposition evinces the point well, I think. The central question there is whether or not simple mono-morphemic words like *house*, *bachelor*, *kill* etc. have non-simple semantics or even compositional semantics. There is now a wealth of evidence in favour of such analyses of words, for adverbials (Morgan 1969), for verbs (Dowty 1979) and more generally Jackendoff (1990) and

Pustejovsky (1995), although the latter's view rejects exhaustive decomposition in favour of certain representational mappings.

The point is that words, often assumed to be the simplest meaningful parts in the PoC, might have meaningful parts of their own. They might have internal structures upon which certain semantic operations can be computed, as the lexical decomposition literature argues. Words exhibit state compositionality on these views. This doesn't mean that when we compute the meaning of sentences we do so via the further compositional structure of words. Words might be state compositional, i.e. decomposable, but the sentences of which they are composed need not be process compositional in terms of them, i.e. the meaning of whole expressions are not built up from the decomposed words.[11]

Pelletier (2012) comes closest to a similar distinction between his "building block version of compositionality" and "functional version of compositionality".

> A difference between the two notions of compositionality concerns whether some 'whole' can contain things not in the parts. According to the building-block view, no; but according to the functional version, yes. For, the first notion allows the whole to contain only what is in the parts, possibly re-arranged in some manner. But the second allows the thing associated with a whole (in the linguistic case: the meaning of a complex whole) to be a function of the things associated with the parts (in the linguistic case: a function of the meanings of the syntactic parts and syntactic mode of combination) (151).

For example, consider the case of neuronal activity in which assemblage-1 of neurons is active during task-1 and assemblage-2 is active during task-2, discussed by Pelletier. Now imagine there is a new task that takes task-1 and task-2 as parts. The question is then whether the assemblage of neurons involved in this amalgamated task is composed of only those neurons involved in either task-1 or task-2? Or does it involve a completely new assemblage of neurons? Pelletier claims that "[i]n the case of describing the neurons active in the complex task, the function *f* need not pick out any of the neurons that are active in the subtasks...but it would still be compositional" (2012: 151) according to the functional account or what I have called state compositionality. My view of state compositionality is broader than this and subsumes the functional approach. Process compositionality then involves recombining or processing the actual materials as they are found in the parts of the computation.

There is a another closely related concept of compositionality which I will call *Outcome Compositionality*. This is a functional notion. It states that given a certain input, the resulting output is compositional. Again, like with state compositionality, the process or input might not be compositional in order for the output to be. The

[11] It might help to think of storage here. Words might be stored as units and brought up or recalled during composition independently of their internal structures. According to Baggio et al. (2012: 656) "psychologically speaking, the real issue is about 'the balance between storage and computation', and the role compositionality plays there". Martin and Baggio (2019: 1) even suggest that "human behaviour, including language use and linguistic data, indicates that composing parts into complex structures does not threaten the existence of constituent parts as independent units in the system: parts and wholes exist simultaneously yet independently from one another in the mind and brain."

difference is that outcome compositionality also need not involve the computation of "meaningful parts" as per the PoC. Any parts will do. An algorithm that operates on pure uninterpreted symbols like a propositional calculus could still generate an outcome compositional structure. It does not need to compute the whole in terms of some property the semantic values of its parts. Therefore it does not entail process compositionality.

Consider a machine translating algorithm that takes idioms as input and outputs compositional sentences. So if the input is the sentence *John kicked the bucket yesterday* the output will be *John died yesterday*. The precise mechanism could involve a simple substitution of terms. The point is that the outcome of such a procedure would be compositional but the input would involve non-compositional meaning. Of course, this would entail state compositionality as well, because outcome compositionality is a special case of state compositionality, namely the local case. Outcome compositionality involves homing in on segments of the whole output and identifying the property of state compositionality at that level.

However, outcome compositionality does not need to involve completed expressions. It can be piecemeal or incrementally evaluated. Thus it does not entail state compositionality (as individual segments can be compositional without the whole output achieving that state). This is due to the fact that "[e]ach step in an algorithm can often be broken down into further sub-steps. We can talk about the algorithm for the whole task [...] or the algorithm for its each of its sub-steps, the sub-steps of its sub-steps, and so on" (Sullivan 2019: 13). This means that compositional structure can be isolated in a piecemeal fashion. Aspects of the construction or expression can be compositional without the entire whole following suit [many of the tools used in machine learning isolate phrase level or even word-pair compositionality such as the SCAN experiments in Lake and Baroni (2018)]. The precise function or algorithm might even change or update during the computation or simulation generating different or branching possible final states. Furthermore, as we will see, these sub-steps might be "black-boxed" or outside of the modeller's comprehension.

Adherents of Marr's famous tripartite analysis of informational systems might put the point in terms of outcome compositionality is at the level of the algorithm while state compositionality is at the level of the computation (and by extension process compositionality would be at the implementation level). Recall that the level of computation asks "What is the goal of the computation, why is it appropriate, and what is the logic of the strategy by which it can be carried out?" (Marr 1982: 25) Here, the goal is semantic composition or decomposition into meaningful parts. But as we have seen, this goal is multiply realisable (as Marr assumes of higher level analyses generally). Contrast this with the algorithmic level where we ask "what is the representation for the input and output, and what is the algorithm for the transformation?" (Marr 1982: 25). Here we are interested in the general procedure used in the computation. The puzzle will later be generated by observing that aspects of the algorithm may not always be transparent in the case of machines (as we will see in Sect. 4).[12]

The above discussion should suffice to provide some intuitive grounding of three important distinctions the nature of which I think is largely neglected in the compositionality literature. In the next section, I shift focus for a moment to the issue of epistemic opacity in AI more generally.

3 Epistemic Opacity

Some philosophers have questioned whether the age of computer science has created any special issues in the philosophy of science and philosophy more generally (Stockler 2000; Frigg and Reiss 2009; Humphreys 2009). Those who claim that it hasn't, often argue that "[t]he philosophical problems that do come up in connection with [computer] simulations are not specific to simulations and most of them are variants of problems that have been discussed in other contexts before" (Frigg and Reiss 2009: 593). They isolate four distinct areas in which computer science is said to present new problems in philosophy. I will focus here on the epistemic variant, as it has been at the centre of more recent debates in the philosophy of AI and AI policy in general.

The issue of trust in AI has garnered public significance recently, especially after the AlphaGo programme defeated the strongest Go players humanity had to offer in 2016.[13] The idea that machines might be capable of "understanding" and solving problems based on experience and learning as opposed to explicit programming instructions of the GOFAI models sent shock-waves through public landscape. Soon policy documents were drafted offering guidelines for "responsible AI" by not only organisations (such as http://www.itechlaw.org/ResponsibleAI and https://ai.google/responsibilities/responsible-ai-practices/) but also political entities (https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai). The central worry seems to be what Humphreys (2009) calls the "anthropocentric predicament" or the problem of "how we, as humans, can understand and evaluate computationally based scientific methods that transcend our own abilities" (617). Nowhere is this alleged predicament more strongly felt than in the deep learning community.

Although Geoffrey Hinton's "deep learning" programme might have ushered in a new era of AI (see Krizhevsky et al. 2012; LeCun et al. 2015), the current state of the field had precursors in much earlier work on connectionism and parallel distributed

[12] There is a tendency in the classical connectionist and current machine learning literature to take compositionality to only involve a recursive relationship between primitive and compound types of some kind (van Gelder 1990, 1994; Baroni 2019). The ways in which this abstract procedure is instantiated are then the particular types of compositionality which are implemented. I think these kinds of definitions run the risk of confusing semantic compositionality with computability and/or combinatoriality. One major difference between the latter concepts and the former is that they can operate on pure strings or syntax without semantic representation. Some experiments in machine learning adopt this confusion and test for compositionality on nonce words or ungrammatical strings. However, the PoC is a semantic principle which is essentially bound up in the syntax-semantics interface and discussions which neglect this aspect can therefore fail to capture its nature.

[13] For comparisons between AlphaGo and Deep Blue of the previous AI generation, see Schubbach (2019).

processing (Rumelhart et al. 1986; Smolensky 1990; Elman 1991). These earlier attempts at modelling human abilities on artificial neural networks met with general philosophical skepticism (Fodor and Pylyshyn 1988; Marcus 2018) largely centred around their limited success especially on so-called symbolic or logical tasks. The recent boom has, in contrast, proven immensely successful on various tasks from image recognition to machine-translation.

Humphrey's proffers the concept of *epistemic opacity* to capture the idea that the abilities of machines might outstrip that of human cognition. The idea behind epistemic opacity is that aspects relevant for knowing or justifying the steps of a particular computation are unknown (or unknowable) by agents of a particular makeup. He claims that "no human can examine and justify every element of the computational processes that produce the output of a computer simulation or other artifacts of computational science" (Humphreys 2009: 618). This might be true in some sense but it overgeneralises. For instance, any mathematical proof based on infinitary logic (Hilbert type or otherwise) would be similarly epistemically opaque (perhaps welcomed news to the Intuitionists). It cannot just be a matter of complexity. Humphrey's offers agent-based models with emergent macro-features as examples of processes inaccessible to human modellers. Apparently, these features only arise once the simulation is running. However, once this simulation runs, humans are able to characterise many of these features and patterns.[14] Similarly, logic-based programming languages such as Prolog, or those based on determinate algorithms, can be extremely complex. But there is a sense in which their steps are comprehensible to their programmers, especially since they designed them to solve a particular task in a particular way.

Epistemic opacity so defined is also a matter of degree. My simple Python program might be epistemically inaccessible to my grandmother no matter how much instruction and explanation I pour in (she might be a hardline C++ advocate). Similarly the algorithm behind IBM's Deep Blue might be epistemically inaccessible or opaque to me.[15] In other words, the kind of epistemic opacity discussed by Humphreys and others is a matter of degree not kind. Many computer programs in the GOFAI tradition were based on logical reasoning and formalised algorithmic procedures (If-then instructions and the like). These of course could be (and are) complex but they aimed at amplifying human ability, not necessarily transcending its nature.[16]

[14] Take Schelling's famous model of segregation. With a minor preference function (30% satisfaction) and two kinds of agents distributed randomly in a population, a macro-level segregation effect is produced. But this equilibrium is explicable in terms of features of the simulation despite the effect only showing itself after a few generations have been run.

[15] I thank Eduoard Machery for pointing this worry out to me.

[16] Humphrey's does go on to define *essentially epistemic opacity* or "a process is essentially epistemically opaque to X if and only if it is *impossible*, given the nature of X, for X to know all of the epistemically relevant elements of the process" (2009: 650). It is unclear what is meant exactly by "epistemically relevant elements" here. Durán and Formanek (2018) interpret it in terms of some sort of surveyability of steps in finite time. Nevertheless, one worries about the historical applicability of some such definition in times before a particular scientific advance. Surely relativity might have seemed epistemically opaque to Newtonians? The definition assumes we have a clear grasp of the limits of our natures and knowledge.

Newman (2016) provides an alternative interpretation of the evidence usually advocated in favour of epistemic opacity. He focuses on Software Engineering and suggests that any impression of opacity is more likely due to the adoption of bad practices on the part of the modellers. He extends his critique to Lenhard and Winsberg (2010) discussion of "confirmation holism" or the alleged impossibility of locating "the sources of the failure of any complex simulation to match known data, so that it must stand or fall as a whole" (Newman 2016: 258). Unlike the standard "proof-checking" model of computer programming in which a modeller self-scans her code to check for bugs or compiling errors, so-called "Big Data" programmes can often be too complex to allow for such a procedure. Lenhard and Winsberg (2010) offer modelling in climate science as a case study, since these complex models involve characteristics such as "fuzzy modularity" in which different models are used to simulate different aspects of the target system.[17] Here Newton's suggestion for better practices is compelling, especially as a way of keeping track of the different models and mechanisms used for one task. For instance, in arguing against Humphrey's notion of essentially epistemic opaque systems (see footnote 17), Newton suggests *decomposition* as a strategy for managing complexity. He further argues that modularity is a benefit to error detection and reduction in Software design and can mitigate the effects of fuzzy modularity (and "kludging" which involves using bits of recalcitrant tools from predecessor models). This is to say that epistemic opacity might be a contingent phenomenon in these cases and indeed better practices can help reduce its effects or "even promote surveyability" (Newman 2016: 567).[18] However, none of these authors explicitly consider machine learning in their discussions, where I will argue techniques like "decomposition" are especially difficult to utilise. Thus, I hope to show that neural nets might pose a particular problem in terms of epistemic opacity, one that does not ride on the issues with the definition or application of notions such as Humphreys'.[19]

In the next section, I argue that it is only with the advent of Deep Learning and the profusion of applications of Deep Neural Networks (DNNs) that the question of epistemic opacity really takes force. It is onto this puzzle and its consequences for the PoC that we move in the next section.

[17] Weisberg (2007) calls this modelling technique "multiple models idealization".

[18] Again see Duran and Formanek (2018) for a computational version of reliabilism as a tool to capture surveyability and epistemic access in the service of grounding trust in complex systems.

[19] Ananny and Crawford (2016) question the ideal of transparency in computational systems itself. They discuss a number of issues with the ideal and conclude that a larger "sociotechnical" appreciation of the interaction between machines and humans is necessary in order to reconstruct the notion of accountability in computational settings. Robbins (2019) also questions transparency but offers "envelopment" of AI systems as an approach to their uncertainty or opacity, in which we contain or limit their impact on and potential harm to humans.

4 The Puzzle

In this section, I will detail an argument for the conclusion that Deep Neural Networks are indeed epistemically opaque in a particular way which prevents an attribution of process and state compositionality to them. They might still be considered outcome compositional if they generalise in specific way, however. First I will present general features of these networks with some focus on the specific subset of them often used for natural language tasks. Then, I will discuss a puzzle about the applicability of the PoC to these models.

4.1 Neural Nets and Deep Learning

Deep learning and Deep Neural Networks (DNNs), upon which deep learning is based, are systems which incorporate multiple hidden layers of connections and weightings which deliver outputs based either on supervised or unsupervised training sets. The aim of such a network is to generalise beyond the training set to a novel test set. Specifically, LeCun et al. (2015: 438) describe the underlying architecture of deep learning as "a multiplayer stack of simple modules, all (or most) of which are subject to learning, and many of which compute non-linear input-output mappings". The DNNs themselves are composed of an input layer, n hidden layers (if n = 1, the network is not very "deep") and an output layer (as shown in Fig. 1 below).

The lines from the input layer feeding into the first set of hidden units (H1) represent the respective weights of each of the input units. You could label these weights according to predefined preferences through supervised learning (basically, the act of labelling the "correct" values by the modeller). The hidden layers in turn represent the non-linear functions or 'activation functions' which take as inputs the values of each of the previous nodes (and their weights). Eventually, after a number of iterations of the process, where the output of one layer serves as the input for the next, it will terminate in the final output.

As for the general purpose of such a network: "DNNs are designed to learn which weights should be assigned to each feature in order to maximize predictive power and identify patterns in data that are not easily detectible by humans" (Sullivan 2019: 20). They can be designed to play Go (Silver et al. 2016), predict future medical illness based on medical records (Miotto et al. 2016), or guess the next word in a sequence (Goldberg 2017). The key to understanding how these models work is by appreciating that their structure allows them to learn and adapt their own algorithms to a particular task. Thus, the path from x_1 to ^y shown in the diagram is misleading as often a clear path is not traceable in a DNN. For instance, *back-propagation* is a hindrance to such a simple analysis. This is when a model, through stochastic tools such as a Batch Gradient Descent optimization function or something similar, corrects for errors of previous epochs (say, from H1 to H2) by fine-tuning the weights in terms of the base error rate.

The neural analogy dates back to idealisations of Connectionism which describe neural nets as idealised and simplified models of real neuronal connections in the brain, with the strength of connections represented via their relative weightings. However, Goodfellow et al. (2016: 16) caution against interpreting deep learning similarly, "one should not view deep learning as an attempt to simulate the brain. Modern deep learning draws inspiration from many fields". Of course, this might complicate matters as it opens the door to analyses like Lenhard and Winsberg's (see above) based on fuzzy modularity and kludging (and perhaps Newton's objections to their conclusions concerning epistemic opacity). But there are dissimilarities between climate science which utilises different models (with different determinate structure) and learning systems such as DNNs in which "the result [of the structure described above] produces a DNN model that follows its own algorithm that it learned through the modelling process" (Sullivan 2019: 23).

There are a number of varieties of DNNs which incorporate architectural features designed for specific types of tasks. For instance, Convoluted Neural Networks work well in image capturing tasks or tasks involving spatial relations. These are feed-forward networks which are non-cyclic and thus unidirectional. They were some of the first kinds of artificial networks and they proved greatly successful on some major applications (LeCun et al. 2015). They differ from Recurrent Neural Networks (RNNs) which are the ones mostly used in natural language tasks, or sequence tasks.

> [A] sequence-processing recurrent network reads some input (e.g. a word), and produces an output (e.g. a guess about the next work) at each time step. The output of the network at time *t* is a non-linear function of the input at time *t*, as well as of the state of the network itself at step t − 1 (weighted by *recurrent connections* that propagate activations across time) (Baroni 2019: 5).

Again, there are various forms of RNNs. For instance, sequence-to-sequence tasks, such as translation between sentences in different languages, are handled well by *encoder-decoder* models (Sutskever et al. 2014). While *gated* RNNs, using Long-short term memory networks or LSTMs, allow for decision procedures to dictate the control of information over time in the network (see Hochreiter and Schmidhuber 1997). Of course, LSTMs and RNNs can also be used as units in encoder-decoder models. Further details of specific architectures are, however, beyond the current scope, the remaining discussion will therefore focus on the idea of black-boxing in DNNs relevant to the present discussion.

4.2 Black-Boxes and Meaningful Parts

Despite the immense success of deep learning on a range of applications, these systems do seem to face a philosophical difficulty, one that is often recognised by their practitioners. Given the architectural designs mentioned above and the self-update and correcting nature of DNNs, aspects of the system can be "black-boxed" from the modeller. The situation is often described in dire terms and is already reflected in the simple stages of the system described in Sect. 4.1:

> [T]he outputs of learning networks are not based on well-defined procedures or explicit criteria any more than their processing. Although we do get an output, we do neither know how this output was computed nor why it is this out-

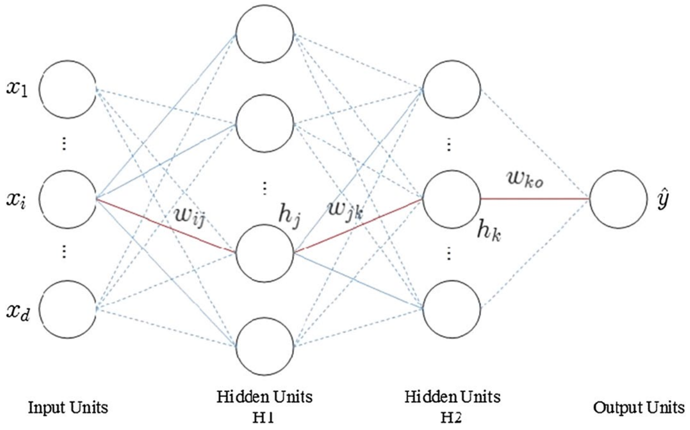

Fig. 1 An example Deep Neural Network

> put and no other. Therefore, DLNs [Deep Learning Networks] are regularly called 'black boxes' (Schubbach 2019: 8).

> In fact the programs are evolving, so when new data comes in, or new feedback is given...the patterns in the learning system change. What this means is that the outcome cannot really be explained, it is not transparent to the user or programmers, it is opaque (Müller 2019: 7).

> The modeller cannot predict which data points will be most salient, nor can the modeller interpret the ways in which the machine settled on certain weights for certain pieces of data...The modeller does not even know which weights or activations will be deployed in a given iteration (Sullivan 2019: 24).

There are some subtle differences in the positions described above. Schubbach is highlighting the differences between DNNs and the determinate algorithmic structures of GOFAI programs, Müller is emphasising the evolving or dynamic nature of these systems in generating opacity[20] and Sullivan is focusing on the possible implementation level opacity.[21] However, there is one thing that most descriptions of black-boxing in DNNs have in common, uncertainty about which parts the system takes to be meaningful or significant.

[20] Technically, dynamics or updates should not preclude the possibility of transparency. Dynamic semantics based as it is on dynamic logic is not epistemically opaque in any sense relevant here and although static concepts of meaning are jettisoned for *context change potentials* or updates, meaningful parts are clearly identifiable. See Groenendjik and Stokhof (1990) and Veltman (1991) for clear descriptions of the general framework.

[21] Sullivan interprets this situation as one of "link uncertainty" in which understanding the intricacies of model is not paramount but rather the epistemic opacity is generated by a lack of understanding the link between the model and target phenomenon.

Let me be more clear. The issue is not that the parts of the learning system are in general opaque. On the contrary, the basic architecture is well-defined and the (statistical) functions used in its operation are decided by the modellers in advance. The issue is not even that we have no means of identifying the inner workings of these systems as is sometimes implied by modellers themselves such as Dudley, as reported by Knight (2017), on the deep patient model in saying "we can build these models, but we don't know how they work".[22] Rather the issue is that we don't know what the machine takes to be a *meaningful part* in its operations. This problem is two-fold. Firstly, it arises at the level of input. It is unclear which parts of the input the machine takes to be meaningful in some cases. Then, this problem is structurally compounded in that the input to further layers is also obscured. Thus, the meaningful parts of the network itself is left opaque.

For example, a particular RNN might do well in generalising to patterns in the test set on a natural language task such as sequence processing. In fact "neural networks display a preference for the grammatical sentences that is well above chance level and competitive baselines" (Baroni 2019: 7) and in some cases approximate the Gold Standard or human level performance (see Gulordava et al. 2018). But how it performs these tasks is still uncertain. In most cases, the indications are of an indirect nature, "it is only through indirect means that the modeller can investigate whether the model is picking up on what seems to be the most relevant features for the task at hand" (Sullivan 2019: 25). The heuristics the system is using might be far from rule-based grammatical constructions. Lake and Baroni (2018) aim to set up benchmarks for compositionality in machine learning by testing the networks on compositional tasks. But the networks failed on tasks requiring systematic compositional rules. They concluded from these experiments that the networks are not compositional yet highly proficient at language processing beyond pattern recognition to structure dependence. An alternative interpretation might be that the much discussed notion of systematicity is not a general feature of linguistic behaviour [see Johnson (2004) for a convincing argument as to the restrictions of systematicity] nor a clear indicator of compositionality [see Werning (2005) for a formal argument] nor indeed is compositionality sufficient for systematicity [see Blutner et al. (2004) for a connectionist argument to this effect]. Nevertheless, the indirect methods used to identify internal processes often rely on the behaviour of the system. Specifically, in the experiment discussed above, outcome compositionality (or its failure) is used to determined process compositionality. But these two kinds of compositionality do not entail one another (as shown in Sect. 2.3).[23]

As previously discussed, there are two essential aspects of the PoC, repeated below:

[22] Many ethical discussions have centred around the possibility or necessity of "opening the black-boxes" or the "right to explanation" (such as the EU's *General Data Protection Regulation* legislation). These discussions are of course beyond the present scope but see Robbins (2019) for an alternative approach to the ethical issues around black boxes in AI.

[23] Similarly, for the salience based methods of describing image classifier tasks discussed in Ribeiro et al. (2016).

> *Meaningful Parts Principle* (MPP): For the meaning of the whole to be determined by the meaning of its parts (and their syntactic combination), there needs to be meaningful parts.
>
> *The Knowable Parts Principle* (KPP): In order to know the meaning of a whole expression, we must be able to identify what the meaningful parts are.

The main problem is that neither principle is respected in the case of DNNs applied to natural language. As we gleaned from the Sullivan quote above, modeller's "cannot predict which data points will be most salient" or how the machine settled on relevant weights i.e. both requirements for identifying parts. Plebe and Grasso (2019) even suggest that it is an open question as to why deep models (with multiple layers) are more successful than shallow models with clear part structure, at least in terms of the mathematics involved. According to my view, the black-boxing specifically obscures the identification of the parts of the process which are considered to be meaningful by the learning algorithm. Indirect methods only get us some traction on outcome compositionality which as we have seen is no guarantee of process or state compositionality. However, if the meaningful parts themselves are black-boxed, then this would make the claim that DNNs or RNNs applied to natural language are compositional (or not) potentially impossible to evaluate. They may indeed turn out to be using compositional rules in their constructions but the epistemic opacity of the system precisely prevents such an analysis.

Recall that process compositionality requires the MPP (and KPP) for an account of how relatively static meaningful parts combine to form larger meaningful parts and eventually entire expressions. Back-propagation and a learning algorithm which through multiple (sometimes tens of thousands) of weighted connections can reorganise itself and reassign weights to its parts makes the isolation of any stable semantic value extremely difficult to understand. The weighting of connections is important here. If we interpret a weighting as a measure of the importance a system assigns to a piece of information or unit in a process (or more precisely between the connections of two properties of the system), then the possibility of that assignment altering significantly without clear explanation creates a black-box effect around that particular unit. If the system is using non-discrete heuristics and non-compositional rules (as perhaps even humans might do) then compositionality of this kind will fail. The problem is that it is not possible to tell.

State compositionality requires decomposition into meaningful parts and, again via a weakened version of KPP, the ability to identify what those parts are, even in the absence of process compositionality. This too is missing in the case of DNNs which do not decompose homogeneously into separable units. We could possibly stop a simulation and "look" at the state of the computation at that stage but given back-propagation, we would still not have a clear path to the final state of the system or the parts that will be relevant to getting there.

What I surmise is that most of the machine learning literature is tracking outcome compositionality. This is what performance on tasks in the test set would be an indicator of, if anything. But outcome compositionality does not entail state or process compositionality, as shown in Sect. 2.3.

Notice, this situation might not be problematic for other applications of DNNs. Visual perception and image recognition tasks are not necessarily compositional (even in humans), so the question of meaningful parts does not obviously come up in that context. On the contrary, I have explicitly argued that the PoC depends on something like the MPP and KPP (above). Thus, when asking the question of whether machines operate compositionally, it matters whether we can identify what they take to be meaningful parts of whole expressions.[24]

In a sense, the situation is not altogether unsurprising. The concept of compositionality originated in applications of formal languages to natural language semantics. The same mathematical underpinnings presupposed in this latter application informed much of the GOFAI approaches to AI. When the mathematics changed with deep learning, this formal property of compositionality can no longer be taken as a given. Groenendijk and Stokhof (2005) go further to claim that since this property is methodologically inherited from formal languages of a certain type (namely, first order), it could have a different or no role in frameworks which do not involve formal languages of the logical variety.

Nevertheless, despite the growing interest in the topic of compositionality in machine learning (Liang and Potts 2015; Baroni 2019; Hupkes et al. 2019; Andreas 2019 to name a few), the kinds of semantic compositionality which have most interested linguists and philosophers of language might not be accessible given our current understanding of artificial networks. Furthermore, the focus on so-called "compositional" solutions to particular learning tasks misses the mark in the absence of identifying meaningful part-structure. In the last section, I will suggest a different potential strategy for discussing the PoC and denuding black-boxes in the context of natural language applications of machine learning.

4.3 A Partial Suggestion

The uniqueness of learning systems is often touted as an obstacle to their comparison. The deep patient model has little in common with the Alpha Go program since the training data vastly differs (medical records vs past Go games), the output differs (patient representations vs 'moves'), the internal structures have different features (unsupervised vs supervised) and of course the tasks were distinct. DNNs can even differ in architecture as shown in Sect. 4.1. The effect of this uniqueness assumption is that theorists and modellers only look at the specific models for indirect evidence as to their inner workings without considering other similar and even dissimilar models.

Here, I suggest that a dual approach might be advantageous in order to approach the black-box issue in deep learning and the compositionality debate in particular.

[24] Of course, compositionality could apply in the visual domain similarly. The argument could go as follows: people seem to interpret visual stimuli they have never encountered before and they do so in a systematic way; the best explanation is that they accomplish this by relying on the smallest interpretable parts of the stimuli and the way those parts are combined. So, visual interpretation must be compositional. I thank Zoltán Szabó for this observation.

Salience mappings and the like might be good tools for homing in on the processes of particular models but cross-model comparisons on similar tasks could identify invariant structure in DNNs more generally.

A proposal in a similar spirit is made in Johnson (2015) with relation to formal grammars in linguistics. The insight is borrowed from physics in which an invariant is a property of a system which remains unchanged under transformation. "Two theories may *prima facie* appear drastically different, and yet be indistinguishable in terms of their empirical predictions, etc. In such a case, they are not essentially different and may be assumed to each capture one and the same underlying idea, albeit in distinct vocabularies" (Johnson 2015: 163). By "indistinguishable in terms of empirical predictions", he means they are notational variants of one another or *weakly equivalent* in the terminology of formal language theory. He further suggests that this idea of notational variants, often inimical to linguists, might point to underlying structural overlap among formalisms and serve to identify the true content of theories "often not identifiable without recourse to notational variants (i.e. symmetries)" (Johnson 2015: 164).

There are a number of extant tools which might be useful for the task of identifying invariant structure. One such tool, borrowed from neuroscience, is ablation. Basically, the idea is that modellers remove some component of the model and measure the effect on the system by comparing the system before and after the removal. In a sense, these studies isolate the behaviour of a system before and after a transformation of sorts. For example, Meyes et al. (2019) perform ablation studies on two dissimilar networks in the computer vision domain showing not only that such studies are useful for identifying structure but also that networks can recover from the removal of components (proportional to the size of the removal in some cases). However, network recovery (which they show can happen after one epoch) can hamper our ability to isolate meaningful parts in a way that is not generally the case in aphasiology.

For structural invariance to be useful at all, it might be necessary to train different DNNs on the same corpus or training set (as is done in computational linguistics with the Wall Street Journal corpus). Similarly predictions might then yield indications of invariant processes among different networks. This approach would mark a shift in the focus from outcome behavioural diagnostics often used to analyse DNNs at present. But there is no reason to avoid using the two approaches in tandem. Of course, different networks might perform differently on the same training sets. But the variation in this case could be illuminating for a comparative account of meaningful part segmentation across different DNNs, i.e. what about architecture A as compared to architecture B resulted in a different output to the same input X.[25]

We are still a ways off from isolating the meaningful parts of such networks but with a dual approach like the one gestured at here, perhaps we can hope to identify

[25] More direct approaches to identifying structure in networks do exist. One famous example is Smolensky's (1990) tensor product representations which aimed at capturing variable binding and symbolic processing while remaining true to the neural net architecture of classical connectionism. See McCoy et al. (2019) for a more recent adaptation of this idea on RNNs.

some invariant structure across black boxes as opposed to only trying to look inside them.

The literature on the black-boxing problem and interpretability is considerably more developed than might be suggested here. Promising research in this vein includes Lei et al. (2016) using the extraction of unsupervised input text as justification (or "rationales") for prediction in neural networks based on a multi-aspect sentiment analysis, and building on this the work of Yu et al. (2019) aimed at identifying corresponding rationales in text matching tasks by means of a three part "rationale generator" using bi-directional LSTM encoder networks. Bastings et al. (2019) who among other things introduce to the former a novel distribution they call a "hard Kumaraswamy distribution" which exhibits both continuous and discrete behaviour, also falls within this approach to the black box problem. In a sense, this kind of research forces the neural models to "show their work" (without sacrificing empirical success) which might provide additional clues as to what DNNs identify as meaningful parts.[26]

This small section merely describes a partial suggestion towards a partial solution to the problem of black-boxes in networks aimed at natural language processing. Much more work needs to be done (and is being done) in order to understand the workings of these extremely successful machines and how they might demarcate the meaningful parts of expressions in order to predict, translate, and learn natural language structures. The lessons learned might in turn hold insights for DNNs designed for other tasks.[27]

5 Conclusion

In this paper, I have harnessed the literature on epistemic opacity in AI to describe a novel puzzle related to artificial networks such as DNNs and the principle of compositionality in linguistics. I have argued that the current state of understanding these networks precludes the possibility of identifying process and state compositional structures within them, which I defined earlier. Finally, I have provided some hints as to how one might begin to find ways of uncovering the black boxes inherent in the neural nets designed for natural language tasks but the puzzle remains for the present state of the science.

Acknowledgements I would like to thank Kyle Blumberg, Eduoard Machery, Zoltán Szabó, Bernhard Weiss and two anonymous reviewers of this journal for their detailed and useful comments on the content. I would also like to thank the organisers and fellow participants of the Compositionality in Brains and Machines workshop held at the Lorentz Center at the University of Leiden in August 2019, especially Marco Baroni, Dieuwke Hupkes, and Jelle Zuidema and the staff of the Lorentz Center itself. Lastly, I would like to thank fellow participants at the Language, Concepts, and Science workshop held at the University of Johannesburg in October 2019.

[26] I thank an anonymous reviewer for pointing me in the direction of this research.

[27] There have been some interesting comparisons between Alpha Go and Alpha Go Zero (which was not trained on human data). See Silver et al. (2017a, b) for more game comparisons.

References

Ananny, M., & Crawford, K. (2016). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, *20*(3), 973–989.

Andreas, J. (2019). Measuring compositionality in representation learning. *ICLR*.

Baggio, G., van Lambalgen, M., & Hagoort, P. (2012). The processing consequences of compositionality. In M. Werning, W. Hinzen, & E. Machery (Eds.), *The Oxford handbook of compositionality* (pp. 655–672). Oxford: Oxford University Press.

Barker, C., & Jacobson, P. (Eds.). (2007). *Direct compositionality*. Oxford: Oxford University Press.

Baroni, M. (2019). Linguistic generalization and compositionality in modern artificial neural networks. Retrieved from ArXiv preprint arXiv:1904.00157, to appear in *the Philosophical Transactions of the Royal Society B*.

Bastings, J., Aziz, W., & Titov, I. (2019). *Interpretable neural predictions with differentiable binary variables*. Retrieved from arXiv:1905.08160.pdf.

Blutner, R., Hendriks, P., De Hoop, H., & Schwartz, O. (2004). When compositionality fails to predict systematicity. In S. D. Levy, & R. Gayler (eds.), *Compositional connectionism in cognitive science. papers from the AAAI fall symposium* (pp. 6–11). Arlington: The AAAI Press.

Brandom, R. (1994). *Making it explicit*. Harvard: Harvard University Press.

Brandom, R. (2007). Inferentialism and some of its challenges. *Philosophy and Phenomenological Research*, *74*(3), 651–676.

Chomsky, N. (1982). *Some concepts and consequences of the theory of government and binding*. Cambridge: MIT Press.

Cooper, R. (1975). *Montague's semantic theory and transformational syntax*. Ph.D. Thesis, University of Massachusetts, Amherst.

Croft, W. (2001). *Radical construction grammar*. Oxford: Oxford University Press.

Davidson, D. (1967). *Inquiries into truth and interpretation: Philosophical essays*. Oxford: Oxford Clarendon Press.

Dever, J. (1999). Compositionality as methodology. *Linguistics and Philosophy*, *22*(3), 311–326.

Dever, J. (2012). Compositionality. In *The Routledge handbook to the philosophy of language* (pp. 91–102).

Dowty, D. (1979). *Word meaning and montague grammar: The semantics of verbs and times in generative semantics and in Montague's PTQ*. Dordrecht: Reidel.

Dowty, D. (2007). Compositionality as an empirical problem. In C. Barker & P. Jacobson (Eds.), *Direct compositionality* (pp. 23–101). Oxford: Oxford University Press.

Durán, J., & Formanek, N. (2018). Grounds for trust: Essential epistemic opacity and computational reliabilism. *Minds and Machines*, *28*, 645–666.

Elman, J. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, *7*, 195–225.

Evans, G. (1981). Semantic theory and tacit knowledge. *Collected papers* (pp. 322–342). Oxford: Clarendon Press.

Fodor, J., & Pylyshyn, Z. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*(1–2), 3–71.

Frege, G. (1908). Über Sinn und Bedeutung. *Zeitschrift fir Philosophie und philosophische Kritik* 100 (1892) 25–50; translated as 'On Sense and Reference' in P. T. Geach and M. Black, Translations from the Philosophical Writings of Gottlob Frege, Blackwell, Oxford, 1960.

Frege, G. (1919). Notes for Ludwig Darmstaedter (Logik in der Mathematik), in Frege 1979: 253–257.

Frigg, R., & Reiss, J. (2009). The philosophy of simulation: Hot new issues or same old stew? *Synthese*, *169*, 593–613.

Fodor, J. (1983). *The modularity of mind*. Cambridge: MIT Press.

Goldberg, A. (2015). Compositionality. In N. Reimer (Ed.), *The Routledge handbook of semantics* (pp. 419–433). London: Routledge.

Goldberg, Y. (2017). *Neural network methods for natural language processing*. San Francisco: Morgan & Claypool.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. Cambridge: MIT Press.

Groenendijk, J., & Stokhof, M. (1990). Dynamic Montague grammar. In L. Kalman & L. Polos (Eds.), *Papers from the second symposium on logic and language* (pp. 3–48). Akademiai Kiadoo: Budapest.

Groenendijk, J., & Stokhof, M. (2005). Why compositionality? In G. Carlson & J. Pelletier (Eds.), *Reference and quantification: The partee effect* (pp. 83–106). Stanford: CSLI Press.

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T. & Baroni, M. (2018). Colorless green recurrent networks dream hierarchically. In *Proceedings of NAACL*, pp 1195–1205, New Orleans, LA.

Haspelmath, M. (2011). The indeterminacy of word segmentation and the nature of morphology and syntax. *Folia Linguistica*, *45*(1), 31–80.

Heim, I., & Kratzer, A. (1998). *Semantics in generative grammar*. Oxford: Blackwell.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Hodges, W. (2012). Formalizing the relationship between meaning and syntax. In M. Werning, W. Hinzen, & E. Machery (Eds.), *The oxford handbook of compositionality* (pp. 245–261). Oxford: Oxford University Press.

Humphreys, P. (2009). The philosophical novelty of computer simulation methods. *Synthese*, *169*, 615–626.

Hupkes, D., Dankers, V., Mul, M., Bruni, E. (2019). The compositionality of neural networks: Integrating symbolism and connectionism. Retrieved from arXiv:1908.08351.

Jackendoff, R. (1990). *Semantic structures*. Cambridge: MIT Press.

Jackendoff, R. (2002). *The foundations of language: Brain, meaning, grammar, evolution*. Oxford: Oxford University Press.

Jacobson, P. (2002). The (dis)organization of the grammar: 25 years. *Linguistics and Philosophy*, *25*, 601–26.

Jacobson, R. (1958/1984). Morphological observations on Slavic declension (the structure of Russian case forms). In L. R. Waugh & M. Halle (eds.), *Roman Jakobson. Russian and Slavic grammar: Studies 1931–1981* (pp. 105–133). Berlin: Mouton de Gruyter.

Janssen, T. (1997). Compositionality. In J. van Benthem & A. ter Meulen (Eds.), *Handbook of logic and language* (pp. 417–473). Amsterdam: Elsevier Science.

Janssen, T. (2012). Compositionality: Its historic context. In M. Werning, W. Hinzen & E. Machery (eds.) (pp. 19–46).

Johnson, K. (2004). On the systematicity of language and thought. *Journal of Philosophy*, *101*, 111–139.

Johnson, K. (2015). Notational variants and invariance in linguistics. *Mind and Language*, *30*(2), 162–186.

Kay, P., & Michaelis, L. (2011). Constructional meaning and compositionality. In C. Maienborn, K. von Heusinger, & P. Portner (Eds.), *Semantics: An international handbook of natural language meaning*. Berlin: Mouton de Gruyter.

Knight, W. (2017). The dark secret at the heart of AI. *MIT Technology Review*. Retrieved from https://www.technologyreview.com/s/604087/the-dark-secret-at-theheart-of-ai/.

Krizhevsky, A., Sutskever, I., & Hinton, G. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* (pp. 1097–1105).

Lake, B., & Baroni, M. (2018). Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *Proceedings of ICML*, pp 2879–2888, Stockholm, Sweden.

Lappin, S. & Zadrozny, W. (2000). Compositionality, synonymy, and the systematic representation of meaning. arXiv:cs/0001006.

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

Lei, T., Barzilay, R., & Jaakkola, T. (2016). Rationalizing neural predictions. In *Proceedings of the 2016 conference on empirical methods in natural language processing*. Association for Computational Linguistics.

Lenhard, J., & Winsberg, E. (2010). Holism, entrenchment, and the future of climate model pluralism. *Studies in History and Philosophy of Modern Physics*, *41*, 253–262.

Leśniewski, S. (1916). Podstawy ogólnej teoryi mnogości. I, Moskow: Prace Polskiego Kola Naukowego w Moskwie, Sekcya matematyczno-przyrodnicza; Eng. trans. by D. I. Barnett: 'Foundations of the General Theory of Sets. I', in S. Leśniewski, Collected Works (ed. by S. J. Surma et al.), Dordrecht: Kluwer, 1992, vol. 1, (pp. 129–173).

Liang, P., & Potts, C. (2015). Bringing machine learning and compositional semantics together. *Annual Reviews of Linguistics*, *1*(1), 355–376.

Marcus, G. (2003). *The algebraic mind*. Cambridge: MIT Press.

Marcus, G. (2018). Deep learning: A critical appraisal. Retrieved from arXiv:1801.00631.

Marr, D. (1982). *Vision*. New York: W.H. Freeman and Company.

Martins, A., & Baggio, G. (2019). Modelling meaning composition from formalism to mechanism. *Philosophical Transactions of the Royal Society B* 375.

McCoy, T., Linzen, T., Dunbar, E., & Smolensky, P. (2019). RNNs implicitly implement tensor product representations. *ICLR*.

Meyes, R., Lu, M., de Puiseau, C. W., & Meisen, T. (2019). Ablation studies in artificial neural networks. CoRR. Retrieved from arXiv:abs/1901.08644.

Miotto, R., Li, L., Kidd, B. A., & Dudley, J. T. (2016). Deep patient: An unsupervised representation to predict the future of patients from the electronic health records. *Scientific Reports*, *6*(26094), 1–10.

Montague, R. (1974). The proper treatment of quantification in ordinary English. *Approaches to natural language* (pp. 221–242). Dordrecht: Springer.

Morgan, J. (1969). On arguing about semantics. *Papers in Linguistics*, *1*, 49–70.

Müller, V. (2019). Ethics of AI and robotics. In E. Zalta (Ed.), *Stanford encyclopedia of philosophy*. Palo Alto: CSLI, Stanford University.

Nefdt, R. (2019). The ontology of words: A structural approach. *Inquiry*, *62*(8), 877–911.

Newman, J. (2016). Epistemic opacity, confirmation holism and technical debt: Computer simulation in the light of empirical software engineering. In F. Gadducci & M. Tavosanis (Eds.), *History and philosophy of computing—third international conference, HaPoC 2015, Pisa, Italy, October 8–11, 2015, Revised Selected Papers* (pp. 256–272). Dordrecht: Springer.

Pagin, P., & Westerstahl, D. (2010). Compositionality I: Definitions and variants. *Philosophy Compass*, *5*(3), 250–264.

Partee, B. (2004). *Compositionality in formal semantics*. Oxford: Blackwell.

Pelletier, J. (2012). Holism and compositionality. In M. Werning, W. Hinzen & E. Machery (eds.) (pp. 149–174).

Pietroski, P. (2018). *Conjoining meanings: Semantics without truth values*. Oxford: Oxford University Press.

Pinker, S. (1984). *Language learnability and language development*. Cambridge: Harvard University Press.

Plebe, A., & Grasso, G. (2019). The unbearable shallow understanding of deep learning. *Minds and Machines*, *29*, 515–553.

Pratt, V. R. (1979). Models of program logics. In *20th Annual Symposium on Foundations of Computer Science (sfcs 1979)*, San Juan, Puerto Rico, USA, pp. 115–122.

Pustejovsky, J. (1995). *The generative lexicon*. Cambridge: The MIT Press.

Pylyshyn, Z. (1984). *Computation and cognition*. Cambridge: MIT Press.

Rambow, O., & Joshi, A. (1992). A formal look at dependency grammars and phrase structure grammars, with special consideration of word-order phenomena. In International workshop on the meaning-text theory. *Darmstadt. Arbeitspapiere der GMD*, *671*, 47–66.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I Trust You? Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144).

Robbins, S. (2019). AI and the path to envelopment: Knowledge as a first step towards the responsible regulation and use of AI-powered machines. *AI & Society*,. https://doi.org/10.1007/s00146-019-00891-1.

Rumelhart, D., McClelland, J., & Research Group, P. D. P. (Eds.). (1986). *Parallel distributed processing: Explorations in the microstructure of cognition: Foundations* (Vol. 1). Cambridge: MIT Press.

Schubbach, A. (2019). Judging machines: Philosophical aspects of deep learning. *Synthese* (online first).

Silver, D., Huang, A., Maddison, C., Guez, A., Sifre, L., van den Driessche, G., et al. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*(7587), 484–489.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., et al. (2017a). Mastering chess and Shogi by self-play with a general reinforcement learning algorithm. Retrieved from arXiv preprint arXiv:1712.01815.

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

Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, *46*(1–2), 159–216.

Stöckler, M. (2000). On modelling and simulations as instruments for the study of complex systems. In M. Carrier (Ed.), *Science at century's end: Philosophical questions on the progress and limits of science*. Pittsburgh: University of Pittsburgh Press.

Sullivan, E. (2019). Understanding from machine learning models. *British Journal of the Philosophy of Science*. (forthcoming).

Sutskever, I., Vinyals, O., & Le, Q. (2014). Sequence to sequence learning with neural networks. In *Proceedings of NIPS* (pp. 3104–3112). Montreal, Canada.

Szabó, Z. (2000). *The Problem of compositionality*. Abingdon: Routledge Press.

Szabó, Z. (2007). Compositionality. In E. Zalta, (ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2007 Edition). Retrieved from http://plato.stanford.edu/archives/spr2007/entries/compositionality/.

Szabó, Z. (2012). The case for compositionality. In M. Werning, W. Hinzen & E. Machery (eds.) (pp. 64–80).

Tarski, A. (1933). The concept of truth in the languages of the deductive sciences. Reprinted in Zygmunt 1995 (pp. 13–172); expanded English translation in Tarski 1983 [1956] (pp. 152–278).

van Gelder, T. (1990). Compositionality: A connectionist variation on a classical theme. *Cognitive Science*, *14*, 355–384.

van Gelder, T. J., & Port, R. (1994). Beyond symbolic: Towards a Kama-Sutra of compositionality. In V. Honavar & L. Uhr (Eds.), *Artificial intelligence and neural networks: Steps toward principled integration* (p. 1071–25). San Diego: Academic Press.

Veltman, F. (1991). Defaults in update semantics. In Hans Kamp (Ed.), *Conditionals, defaults and belief revision*. Dyana Deliverable R2.5A: Edinburgh.

Weisberg, M. (2007). Three kinds of idealization. *The Journal of Philosophy*, *104*(12), 639–659.

Werning, M. (2005). Right and wrong reasons for compositionality. In M. Werning (Ed.), *The Compositionality of Meaning and Content* (vol. 1, Foundational Issues, pp. 285–309). Frankfurt: Ontos Verlag.

Werning, M. (2012). Non-symbolic compositional representation and its neuronal foundation: Towards an emulative semantics. In M. Werning, W. Hinzen, & E. Machery (Eds.), *The Oxford handbook of compositionality* (pp. 633–654). Oxford: Oxford University Press.

Wittgenstein, L. (1953). Philosophical investigations. In G. Anscombe & R. Rhees (Eds.), *G.E.M. Anscombe (trans.)*. Oxford: Blackwell.

Yu, M., Chang, S., & Jaakkola. T. (2019). Learning corresponded rationales for text matching, 2019. Retrieved from https://openreview.net/forum?id=rklQas09tm.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.