Minds and Machines (2023) 33:507–524
https://doi.org/10.1007/s11023-023-09642-0

Can Deep CNNs Avoid Infinite Regress/Circularity in Content Constitution?

Jesse Lopes[1]

Received: 9 March 2022 / Accepted: 4 July 2023 / Published online: 17 July 2023

Abstract

The representations of deep convolutional neural networks (CNNs) are formed from generalizing similarities and abstracting from differences in the manner of the empiricist theory of abstraction (Buckner, Synthese 195:5339–5372, 2018). The empiricist theory of abstraction is well understood to entail infinite regress and circularity in content constitution (Husserl, Logical Investigations. Routledge, 2001). This paper argues these entailments hold a fortiori for deep CNNs. Two theses result: deep CNNs require supplementation by Quine's "apparatus of identity and quantification" in order to (1) achieve concepts, and (2) represent objects, as opposed to "half-entities" corresponding to similarity amalgams (Quine, Quintessence, Cambridge, 2004, p. 107). Similarity amalgams are also called "approximate meaning[s]" (Marcus & Davis, Rebooting AI, Pantheon, 2019, p. 132). Although Husserl inferred the "complete abandonment of the empiricist theory of abstraction" (a fortiori deep CNNs) due to the infinite regress and circularity arguments examined in this paper, I argue that the statistical learning of deep CNNs may be incorporated into a Fodorian hybrid account that supports Quine's "sortal predicates, negation, plurals, identity, pronouns, and quantifiers" which are representationally necessary to overcome the regress/circularity in content constitution and achieve objective (as opposed to similarity-subjective) representation (Burge, Origins of Objectivity. Oxford, 2010, p. 238). I base myself initially on Yoshimi's (New Frontiers in Psychology, 2011) attempt to explain Husserlian phenomenology with neural networks but depart from him due to the arguments and consequently propose a two-system view which converges with Weiskopf's proposal ("Observational Concepts." The Conceptual Mind. MIT, 2015. 223–248).

Keywords Deep learning · Concepts · Object identity · Objective representation · Semantic segmentation · Similarity semantics · Content identity · Language of thought · Phenomenology

ἀδύνατον δὲ ἄπειρα ἐπικρῖναι—Sextus Empiricus, *Outlines of Pyrrhonism* II. 78.

✉ Jesse Lopes
redbroom@bu.edu

[1] Boston College, Boston, USA

In this paper, I argue that deep convolutional neural networks (CNNs) necessarily result in infinite regress and circularity in content constitution if CNNs employ the empiricist theory of abstraction, which results in a similarity semantics. Since the empiricist theory of abstraction has been shown to be employed by deep CNNs (Buckner, 2018), it follows that content constitution in CNNs cannot avoid the infinite regress and circularity mentioned, unless supplemented by (what Quine called) ‘the apparatus of identity and quantification,’ which allows for an identity semantics (as opposed to a similarity semantics). The entailment, its consequences, and the remedy will be the subject of this essay.

1 Infinite Regress & Circularity in Deep CNNs via the Empiricist Theory of Abstraction

In this section, I’ll outline the empiricist theory of abstraction and my relation to it. I’ll then present the arguments against the theory, with reference to the type of encoding found in deep CNNs. The next section discusses deep CNNs in detail and how the arguments apply.

The empiricist theory of abstraction is the idea that general representations, i.e. concepts, might be constituted, with regard to their contents, by noticing sensed similarities among multiple objects and subtracting from their differences. Another way to describe the theory, with a view to the objects that correspond to the concepts thus generated, is that “the empiricist theory of abstraction…. [i]s the theory… that abstract objects arise by our directing attention to some aspects of what we experience and overlooking others: the retained features constitute the abstract object…” (Simons, 1995, p. 107). Consequently, the abstract object (e.g. square, triangle, cat, dog, table, chair, red, green) will be a reflection of amalgamated similarities from sense-experience, with differences subtracted. The object corresponding to a quasi-concept generated in this fashion would be, as Quine would have it, a “half-entity,” corresponding, for Quine’s behavioristic scruples, to a mass noun, or (for us) to an amalgamated content similarity (Quine, 2004, p. 107). To make the object a ‘full entity,’ (an *objectively* meant entity) it would need to correspond representationally to a system instantiating “the apparatus of identity and quantification” (ibid.). For Quine, this was a behavioristically interpreted natural language, with its symbolic predicates and quantifiers. For me, this is a representational theory of the mind, involving the panoply of traditional computational elements and operations: “sortal predicates, negation, plurals, identity, pronouns, and quantifiers” (Burge, 2010, p. 238). But now, since Quine’s criteria for this assumption have been shown to be satisfied by children *prior to* the mastery of a natural language (Carey, 2009), it would seem to follow that what corresponds to a ‘full entity’ (as opposed to a ‘half-entity’) is a *content identity* (as opposed to a content similarity) in a language of thought (LOT). Section III discusses how similarity contents might be compatible with LOT.

How does the empiricist theory of abstraction work, and why are deep CNNs committed to it? The general idea of how the theory works is this: one can acquire the concept RED and the concept TRIANGLE by being minimally exposed to three objects: a red square, a red triangle, and a green triangle. One thereby notices that the sense experience of red is abstractable from the shape, just as the shape of a triangle is abstractable from its color. We thus have a rudimentary picture of how general representations, applicable to many objects, might be abstracted from sense-experience—this being the primary alternative to the doctrine of innate ideas. Deep CNNs are committed to this picture of how the mind arrives at general representations according to Cameron Buckner: "The classical empiricists never specified a plausible mechanism that could perform the crucial 'leaving out' of irrelevant information highlighted in abstraction-as-subtraction. This, I argue, is the role played by max-pooling units" (2018, p. 5357). Having been convinced by this argument, I shall discuss below the way in which max-pooling units implement the empiricist theory of abstraction. For now, one need only note, to see the connection, that max-pooling units 'pool' together multiple features, ignoring differences, to create a general representation, which allows for accurate classification of new inputs *to the degree that they are similar to the old*. As a result, many people believe that deep CNNs show us how one might learn a "concept" and its corresponding "object identity" (Goodfellow, Bengio, Courville, 2016, p. 17). Thus the empiricist theory of abstraction is a theory of content constitution that amalgamates similarities into generalities by a method of abstracting-as-subtracting; and deep CNNs are committed to this theory of content constitution in virtue of their architecture.

In the *Logical Investigations*, Husserl argues (anticipating Fodor) that this theory isn't going to work for concepts (or universal meanings), and their corresponding objects (which are properly neither individuals *nor amalgamations of them*). The reason is simple: concepts require content identity in order to enter into logical relations; what it is *to be* a concept is to be a meaning susceptible of the "apparatus of identity and quantification" (Quine, 2004, p. 107). This is an assumption which is, at least in practice, if not in theory, admitted by neural network theorists. For a meaning must have a unifying structure, capable of recurring and being manipulated as a unit. Today we would call this 'syntax.' Many of Husserl's points on this topic directly anticipate Fodor's and remain explananda for deep neural networks. For example, Husserl begins his criticism of the empiricist theory of abstraction (in the 2nd *Logical Investigation*) by noting that each "meaning certainly counts as a unit in our thought and that on occasion we pass evident judgements upon it as a unit" (2001, p. 241). The evidence of such judgments is a reason to think that the unitary, discrete nature of such meanings is not an artifact of linguistic usage (in the manner of Quine) but rather an indication of genuinely discrete, unified meanings "in our thought." His examples include the analogy to grammatical words: "an identical subject for numerous predicates;" and compositionality: "It can be summed together with other meanings and can be counted as a unit" (ibid.). These examples point to the classical virtues of systematicity and compositionality of symbolic representations in a language of thought (LOT). These remain explananda for deep learning. For Husserl, all of these points draw the same moral: concepts cannot be constituted by empiricistic abstraction; and for us (a fortiori) cannot be constituted by deep CNNs. Thus in virtue of the nature of concepts and the relevant explananda, one begins to think that deep CNNs will be inadequate to the phenomena.

Having set the stage, I'll now present the arguments. Husserl was perhaps the first to note that empiricist theories of abstraction, *a fortiori* deep CNNs, result in infinite regress and circularity in content constitution. This is later echoed insistently and very seriously by Fodor & Lepore (1992) & Fodor (1998, 2008); it is, moreover, at least implicit in Quine's (corpus-wide) distinction between similarity 'encounters' (expressed by mass nouns) and the apparatus of identity and quantification (reflected by count nouns), which Quine believed to derive from one's linguistic community (as opposed to a representational mind) (Quine, 2004). Although we are not here interested precisely in exegesis, it is nevertheless appropriate to look at the exact wording of Husserl's original argument, concerning the infinite regress between content-identity and similarity groupings:

> The conception we are criticizing operates with 'circles of similars' (*Ähnlichkeitskreisen*), but makes too much light of the difficulty that each object belongs to a plurality of 'circles of similars' and that we must be in a position to say what distinguishes these 'circles of similars' among themselves. It is plain that, in default of a previously given Specific Unity, we cannot avoid a regress in infinitum (2001, pp. 243–244).

Content constitution via similarity cannot logically lead to a "specific unity" necessary for meaning. If similarities in the input are being computed at the most basic level of content-constitution, then the next level up, according to Husserl's argument, can only be "similarities of… similarities" (Husserl, 2001, p. 244). In the literature on deep CNNs, this is referred to as "'patterns of patterns'" of similarity groupings (Dube, 2021, p. 76). If these similarities of similarities are grouped once more by the Hubel & Wiesel inspired 'complex cells' of the upper layers of a deep CNN, the next level up, again, will consist of similarities of similarities of similarities. And so on, ad infinitum. This is the infinite regress, which results "in default of a previously given Specific Unity" (2001, p. 244). What Husserl means by this, qua content constitution, is that there is no level of generalization, *starting with* similarities and differences, at which the (content) identity necessary for phenomenologically evident, discretely unified meaning emerges. For similarity is not identity (*simile non est idem*). Thus if meanings *just are* specific unities or discrete units of thought with generally applicable content, the empiricist theory of abstraction, a fortiori deep CNNs, cannot result in meaning but can only "approximate meaning" (Marcus & Davis, 2019, p. 132).

The regress, then, results from not being able to invoke content identity, which would block the regress. A vector space that encodes for red objects, and a vector space that encodes for triangular objects, may overlap in the representation of some object A, a red triangle. But we can only say, given network resources, that this overlapping is similar to similar overlappings, on which the network has previously been trained. Let's say B is a red square and C is a green triangle. Then the argument is this:

1. A is similar to B in one respect (red-likeness) and C in another respect (triangularity-likeness).
2. The respect in which A is similar to B is defined by its similarity to previously observed (inputted) similarities.

3. The respect in which A is similar to C is defined by its similarity to previously observed (inputted) similar similarities.
4. These previously observed (inputted) similar similarities are in turn defined by previously observed (inputted) similar similarities.
5. Premise 4 applies *ad infinitum*, generating illegal species and genera ('illegal' because assuming what was to be proved (*petitio principii*)).

A vector space, let us assume, may encode the content red-triangle-like, and so represent the corresponding similarity bundle object. But where do we find the identical respect around which the similarity grouping is organized? Husserl & Fodor argue that we will be referred to the similarity species (*Art*)—red-triangle-like—and the similarity genus (*Gattung*)—red-like, the 'contents' of which are simply begged in the explanation. That's in part why Fodor & Pylyshn (2015) claim that "connectionists/associationists have no theory of conceptual content" (51). The infinite referral highlighted by Husserl is a failure to respond to the question of how did the network identify the respect in which a group of similar encodings in vector-space is similar without already knowing *the identity in respect of which* the group of similars is similar. If there is no answer to this question, or if the answer is negative, then neural networks, no matter how deep (or accurate), cannot be said to be recognizing objects (per se). Rather, they may be said to recognize, from our conceptually grounded perspective, *approximate* objects (half-entities) or similarity bundles.

The reasoning behind the above infinite regress is simple and points to a fundamental circularity in the empiricist approach. At each step in the infinite regress—from phenomenal features (whisker-*like* content), to species (cat-*like* content), to genera (felis-*like* content)—we "come up against kinds" that are not kind-*like*, i.e., not constituted in terms of continuous similarity spaces (Dube, 2021). At each step in the classificatory hierarchy, therefore, we are already presupposing what we are *therefore circularly* seeking. We "cannot predicate," as Husserl says, "[even] exact likeness of two things, without [already] stating the [identical] respect in which they are thus alike" (2001, p. 242). There is thus a phenomenologically evident asymmetrical dependency between kind-*like* or meaning-*like* representations on kind-*simpliciter* or meaning-*simpliciter* contents. The phenomenology is that when we *mean kinds* we do not mean anything kind-*like*. To explain our intentionality toward kinds on the basis of kind-*like* representations is to presuppose what was to be explained, since one cannot even state the similarity content except in terms of the content-identity. Notice the reverse is not true. One can mean (intend) objects and refer to them *without any dependency on the notion of similarity*. The explanation of concepts and objects qua identity based on similarity is therefore (also) circular.

None other than Fodor in *Concepts* (1998) treads the same ground here as Husserlian phenomenology:

> It looks as though a robust notion of content similarity can't but presuppose a correspondingly robust notion of content identity. Notice that this situation is not symmetrical; the notion of content identity doesn't require a prior notion of content similarity (32).

Fodor is here arguing against a proposal of Gilbert Harman's to the effect that the nature of concepts should be theorized in terms of similarity spaces (a perennial desire of empiricists). If Husserl and Fodor are right, however—and I see no argument to the contrary—concepts can never be explained in terms of any explanatory apparatus which essentially refers content to the output of inductions from similarities (cf. Carey, 2009, p. 28). This is problematic because, as Fodor & Lepore independently argue, "content similarity actually presupposes a solution to (and therefore begs) the question of content identity" (1992, p. 197). Thus the circularity argument looks like this:

1. All concepts exhibit content identity.
2. Empiricism and deep CNNs propose that the contents of all concepts are built (via some similarity metric) from more or less similar (real) individuals.
3. But similarity presupposes identity (Husserl, 2001; Fodor, 1998).
4. Therefore, all content constitution explanations from similarity are circular.

This argument appears to be inescapable, but only if concepts corresponding to "object identity" are explananda for empiricistic deep CNNs (Goodfellow et al., 2016, p. 17). There is some reason to think they are, insofar as the literature speaks freely of concepts in relation to object identity (Goodfellow et al., 2016; Kelleher, 2019; Dube, 2021). Insofar as these 'concepts' are the products of the empiricist theory of abstraction instantiated by deep CNNs, they will be subject to the above arguments. That's a problem if deep CNNs (or generally DNNs) are thought to potentially explain and *model* the tokening of concepts, as has been recently argued (Shea, 2021; Dube, 2021). For (as premise 1 states) there is a "non-negotiable" publicity constraint on concepts on the content-side, which involves content identity (Fodor, 1998, pp. 33–34; Fodor & Pylyshyn, 2015, p. 55; Hopp, 2011). This constraint is violated by deep learning if the above argument is valid. The reason why one might adhere to the constraint is this: one cannot (phenomenologically) intuit a kind *as* a kind, for example, unless one tokens a concept susceptible of an identity semantics—for kinds are not kind-*like*. But, since we cannot "come up against kinds" according to the "empiricistic conception" due to reliance on an unexplicated notion of kind-*like*, it follows that deep CNNs cannot result in the specific unities of conceptual meanings required by a semantics capable of logic (Husserl, 2001 p. 244).

The arguments above apply to content-constitution. There is a lesson for the object-side as well. If deep CNNs cannot result in concepts, they cannot, on the object-side, 'objectively' represent objects. This is certainly a paradox, not least because "simple object recognition is deep learning's forte" (Marcus & Davis, 2019, p. 108). But objective representation, on pain of representing "half-entities, inaccessible to identity," requires discretely unified symbolic identities (Quine, 2004, p. 107). Content identity is qualitatively distinct from amalgamated similarity groupings. Consequently if deep CNNs cannot achieve content identity, it follows they cannot objectively represent *objects*—for "no entity without identity" (Quine, 2004, p. 107). At best, deep CNNs can statistically approximate (without theoretically explaining) conceptual content and the representation of objects through "circles of similars" (Husserl, 2001, p. 243; Marcus & Davis, 2019, p. 132).

The above arguments, specifically aimed at the empiricist theory of abstraction, apply to deep CNNs (by transitivity) (Buckner, 2018). Husserl thought these were knock-down arguments against the theory, necessitating "the complete abandonment of the empiricist theory of abstraction" (2001, p. 114). If that evaluation is correct, these arguments would compel us to likewise 'completely abandon' deep CNNs (by transitivity)—at least as regards their scientific (as opposed to their engineering) interpretation. This paper proposes to take these arguments seriously. It is therefore necessary to look more closely (in the next section) at the machine that has given life to a refuted theory (Buckner, 2018). Instead of completely abandoning the theory, however, as Husserl recommended, the section after the next (section III) will consider how the theory may be salvaged as potentially explaining similarity judgements, without, however, being a theory of concept attainment.

2 Deep CNNs & the Empiricist Theory of Abstraction

In this section, I describe deep CNNs in some detail to show that they employ the empiricist theory of abstraction and therefore are subject to the above arguments.

Deep CNNs were especially designed for image classification tasks. Their "success" in the past ten years has been described as "tremendous" (Goodfellow et al., 2016, p. 321) and "incredible" (Kelleher, 2019, p. 138). They were directly inspired by Hubel & Wiesel's (1962) discovery that neurons in mammalian cortex are specialized to fire in response to proprietary stimuli (e.g. slits, edges, contrast bars). Hubel & Wiesel deemed these neurons 'simple cells,' as opposed to 'complex cells,' which combine input from the simple cells. Fukushima's Neocognitron (1980) applied this idea to neural networks. The key realization was that if a network layer *shares* a set of weights, called "parameter sharing," then the layer's receptive field will be fixed in a manner similar to Hubel & Wiesel's simple cells (Goodfellow, Bengio, Courville, 2016, p. 326). In practice, this means that if a pixel pattern of the 2-D input is present *anywhere* in the image, the function defined by that layer of shared weights, provided that it scans the area with the pattern, will record its presence in the output (a visual feature map). Since the same point applies to the pooling of patterns from several layers, the general representations that result at the pooling layer become "translation invariant" (Kelleher, 2019, p. 168). Translation invariant content is *detached* from the location at which its object was initially recorded. A generalization process thereby occurs. This is why there is great plausibility in the idea that deep CNNs contribute to the (empiricistic) explanation of general representations of the mind.

The basic outline of the initial processing of the machine should be clear. There is a 2-D topological input (or, if the input consists of time-series data, 1-D). There is then a layer of shared weights, known as the 'kernel matrix,' which is the convolutional layer. This layer searches the image for proprietary stimuli, like a flashlight scanning a darkened room for a particular stimulus (Kelleher, 2019, p. 162). And then there's the output of the convolutional layer, which is the visual feature map, recording the presence of various pixel patterns. The reader may note that, since pixel patterns are not themselves representational, the representations generated from them are by definition sub-symbolic (Shea, 2021).

The output of the feature map now becomes the input to the final processing layers. There are typically three of these—a nonlinearity layer, a pooling layer, and a dense layer. So after the visual feature map is generated, it becomes the input to a nonlinear activation function layer, which updates the topological carving of the input space, typically with ReLU (Dube, 2021, p. 68). ReLU is a nonlinear function which changes all negative values to zero. This means that neurons below a certain threshold are cut off entirely from the adjacent pooling layer (Sejnowski, 2018, p. 132).

The pooling layer, therefore, comes next; and this is the operation on the data structure of the visual feature map that is of greatest philosophical interest, since it serves as the focus of Buckner's identification of deep CNNs with the empiricist theory of abstraction (2018). What we wish to argue is that this pooling layer represents a detachable similarity-content (or amalgam), whose denotation is, in the words of Quine, "a half-entit[y], inaccessible to identity" (Quine, 2004, p. 107).

Since our arguments rest on this identification, some definitions from deep learning's practitioners and theorists are in order. "A pooling function," say Goodfellow et al. (2016), "replaces the output of the net at a certain location with a summary statistic of the nearby outputs" (330). This "summary statistic" is an amalgamation of similarities. For this is the locus in the network of the empiricistic 'abstraction-as-subtraction' operation, previously identified by Buckner's argument (2018). What is being subtracted are the dissimilarities; what remains are the similarities. The similarities, therefore, constitute the content of the "summary statistic." And thus, the abstraction (and any content attributable to the machine) is a similarity amalgam.

Now, if this *is* the general representation to be identified with a concept—and this appears to be the general presumption in the field (Goodfellow et al., 2016 *passim*; Buckner, 2018; Kelleher, 2019; Dube, 2021)—then the content of this concept is a similarity amalgam. But since a concept is not a concept unless it is the vehicle of content-*identity* (according to logical considerations), and since there is logically no equivalence between similarity and identity (*simile non est idem*), deep CNNs cannot be said to attain concepts (see section IV for a discussion of the possibility of ignoring the relevant logical considerations).

The problem arises immediately from Buckner's identification of deep CNNs employing the empiricist theory of abstraction, which for the same reasons cannot achieve concepts (or object identity). That's a problem because the empiricist theory of abstraction is a theory of concept attainment. This understanding led Husserl to seriously argue that the empiricist theory (*a fortiori* deep CNNs) must be completely abandoned. CNNs are, in this way, technically refuted as a possible explanation for how the mind achieves "concept[s]… and object identity" (Goodfellow et al., 2016, p. 17).

Our reasoning, however, may still seem too quick for this conclusion. I propose, therefore, to conclude this section by briefly looking at the paper that started 'the deep learning revolution'—the famous AlexNet paper (Krizhevsky, Sutskever, Hinton, 2012).

The paper that first describes the similarity semantics of deep learning—the one that started the deep learning 'revolution'—is "ImageNet Classification with Deep Convolutional Neural Networks" (2012). This paper describes the performance of a supervised convolutional neural network (CNN) on image classification tasks. A CNN is (again) specifically designed for image recognition tasks. The idea here (again) is that the nodes in the early layers extract phenomenal features from the raw pixel values of the input. These features get aggregated or amalgamated in later processing layers of the network. At each stage, these representations are sub-symbolic, since they are constituted from elements that are not themselves representational. The combined input primitives (e.g. oriented lines) form, in later layers, higher-order and more recognizable features (e.g. fineness-of-fur) which, in the final layers, get induced into readily recognizable complexes (e.g. cats). The generalization process occurs naturally by training multiple filters (layers of nodes) to respond specifically to certain pixel values and features, through parameter sharing or "tied weights" (Goodfellow et al., 2016, p. 328). Generally, the convolving filters (or kernel matrices) are several orders of magnitude smaller than the input space, encouraging generalization (Goodfellow et al., 2016, p. 326). By sequentially convolving the filters across the entire input space, the features will be detected if they are present.

The outputs of the filters after the pooling layer can be combined in several ways. The ImageNet CNN used what are called 'dense' layers toward the output. The dense layers are typically the final layers of the network. The layer is 'dense' because, in contrast with the rest of the network, which bears the property of 'sparse connectivity'—meaning not all neurons are connected with all other neurons in the preceding layer—each of the nodes in the dense layer relate to all of the outputs of the preceding pooling (or dense) layer. In this way, the dense layers are more like regular ANNs.

The important philosophical point is that the representational semantics of AlexNet is explicitly a similarity semantics. As a result the above infinite regress/circularity arguments apply. This is brought out in Sect. 6, where the authors consider the dense layers of the network. These final layers have 4096 neurons each. The authors state that one way "to probe the network's visual knowledge" is the following:

> [C]onsider the feature activations induced by an image at the last, 4096-dimensional hidden layer. If two images produce feature activation vectors with a small Euclidean separation, we can say that *the higher levels of the neural network consider them to be similar*. [....] Computing *similarity* by using Euclidean *distance* between two 4096-dimensional, real-valued vectors is inefficient, but it could be made efficient by training an auto-encoder to compress these vectors to short binary codes (Krizhevsky, Sutskever, Hinton, 2012, p. 8 italics added).

Notice that it is the (Euclidean) distance relations which are defining the content for the modelers in terms of "similarity." The content is defined over a continuous region of representational space; it is not discrete (i.e. symbolic). Proximity in representational space defines a geometrical region of similarities, with dissimilars spreading outward. Content is, therefore, understood as a similarity-amalgam—induced similarities deriving from similarities in pixel patterns from the input. We can also call these induced CNN patterns "'patterns of patterns'" or similarities of similarities (Dube, 2021, p. 76). One can begin, at this point, to catch a glimpse of the infinite regress which Husserl calls "the worst of infinite regresses"—a charge originally leveled against the empiricist theory of abstraction *simpliciter*, one which, however, carries over to deep CNNs due to their employment of the empiricist theory of abstraction. For the degree to which any content-similarity, corresponding to an activation pattern, is similar to any other is the degree to which its circle of (induced) similar image patterns is close to the other circle of (induced) similar image patterns.

The appeal to auto-encoders magnifies the issue. Motivation for the suggestion—to use auto-encoders for greater efficiency—explicitly includes creating a tighter correspondence between Euclidean similarity and "semantically similar" representations (review the quote above). Auto-encoders themselves operate by compression (traditionally for feature learning tasks) to give a useful model that "resembles"—by application of abstraction-as-subtraction—the input (Goodfellow et al., 2016, p. 493). The result would be, in the most efficient scenario, representation-building from input similarity through autoencoder similarity to (sub-symbolic) semantic similarity, defined over continuous distance relations (Churchland, 2012, pp. 38–45). The recognition the machine can be said to perform will be a recognition via similarity amalgams of "half-entities, inaccessible to identity" (Quine, 2004, p. 107). Whatever content the machine can be said to build from its input will be on the basis of "computing similarity,“ not identity (Krizhevsky, Sutskever, Hinton, 2012, p. 8).

3 A Hybrid LOT Theory: Quine's Apparatus + Yoshimi's Dynamics qua Husserlian Phenomenology

In this section I try to find a half-way point between the statistical learning of deep CNNs and the logical conditions on concepts and object identity that are essential to a theory of concepts that avoids the infinite regress/circularity of content constitution.

The previous section argued that the "summary statistic" of the max-pooling units of a CNN is a similarity content (of similarities) (Goodfellow et al., 2016, p. 330). The statistical summation of the max-pooling units is a similarity amalgam, to which the arguments of section I apply. It follows that the representations of the network are approximate similarity contents that, if they are to explain concepts and their objects qua identity, result in a disruption of logic (Quine, 2004, p. 107). The result is not peculiar to deep CNNs—it applies just as much to any theory of concept attainment deriving from, and including those of, the British empiricists.

Now one might follow Husserl's startling recommendation that "the empiricist theory of abstraction must be completely abandoned" (2001, p. 114). If so, one should likewise completely abandon the currently popular idea that deep CNNs might illuminate the nature of conceptual tokening. But due to the tremendous success of deep CNNs—they regularly outperform humans on a variety of tasks—we might think this remedy too strong. Since there is one philosopher, Jeffrey Yoshimi, who would particularly recommend this in relation to Husserlian phenomenology, and since we have been guided by the (logical) phenomenology throughout our discussion, we can perhaps begin developing a hybrid model by considering Yoshimi's approach.

Yoshimi has spent the last two decades valiantly attempting to unite the phenomenological descriptions of Husserl with the empiricism of neural networks (e.g. Yoshimi, 2011). As should be clear from our arguments, however, this project will eventually run into the infinite regress and circularity of content constitution. But we might, nevertheless, start with Yoshimi to more clearly show how to unite the two main approaches to cognitive science (Marcus, 2001; Cain, 2016). This should have the corollary of providing a beginning for a complete explanatory (causal) theory for the phenomenology of logical experiences, one that does not "disrupt logic" by reducing discrete symbolic units of thought to continuous gradients (Quine, 2004, p. 107).

The centerpiece of Yoshimi's *Husserlian Phenomenology: A Unifying Interpretation* (2016) consists of two functions that output a continuous gradient (x ∈ [0, 1]). One is an expectation function, the other an update function (17, 29). These functions are intended to generate a similarity semantics of sub-symbolic contents (43). To illustrate how these functions work in experience, Yoshimi chooses the Humean example (from Kant's *Critique of Pure Reason*) of moving around a house with one's body. Through statistical learning, one forms associations (expectations) about house-like aspects of experience (18). The more these associations are confirmed by one's bodily movement, visual experience, and background knowledge, the higher the returned expectation gradient. An update rule supplements the expectation rule by producing "incremental changes in background knowledge" (31).

Provisionally accepting Yoshimi's account as a starting-point, we can hypothesize that the machinery of deep CNNs might supply the values to the visual experience and background knowledge variables. Degrees, therefore, of sub-symbolic visual fulfilment—an expectation being satisfied in accordance with past associations—can thereby be causally explained. Yet we still do not have what would stop circularity and regress "even in the sensory realm," as Husserl says, if we define content identity as "a limiting case of 'alikeness" in accordance with Yoshimi's continuous outputs (2001, p. 242). To stop the regress and prevent circularity, we would need to "come up against kinds" (Husserl, 2001, p. 244). We are, in other words, in search of those specific unities (e.g. HOUSE), for which there are unities of fulfilment (and frustration). Yoshimi has the empiricistic cart before the phenomenological horse; he has explained degrees of fulfilment before he has given any account of what is being fulfilled. What's being fulfilled, according to Husserlian phenomenology, are the meanings corresponding to the objects of logical experience: discrete units of thought corresponding to identical kinds. These kinds, as the objects of an act of knowledge, must be the objects of an identity semantics, as opposed to a similarity semantics.

Without casting Yoshimi and deep CNNs entirely aside, it is necessary to propose a two-system view, given the regress and circularity above, and the remarkable coincidence between the Husserlian and Fodorian projects. This is distinct from single-system views (Marcus, 2001, Smolensky & Legendre, 2006) and coincides with Weiskopf's recent proposal concerning conceptual identity semantics being *qualitatively distinct* from similarity semantics (Weiskopf, 2015, p. 239). Accordingly, there must be a system of statistical learning based on similarity metrics, which may, however, be jettisoned as inessential when considering a system of knowledge in the abstract; and a distinct system based on conceptual units and computational transformations supporting a quantificational syntax and identity semantics (Fodor, 2008, pp. 159–163). The two *may* causally interact but they must be seen, due to regress/circularity, as *in principle distinct*.

We think the two systems may interact via Yoshimi's update rule (or "learning rule λ" see Yoshimi, 2016, p. 31). As the asymptotic approximation converges (its cat-*like*, dog-*like*, house-*like* representations) toward a conceptual kind (CAT, DOG, HOUSE etc.), the representation in the neural network is drawn toward 'a zone of stability.' Such 'zones' can be thought of as attractors in state space. An attractor can be thought of as a set "to which all neighboring trajectories converge" (Strogatz, 2015, p. 331). In this way, there is a brute-causal transition from the one system to the activation of the other.

To bring all this together in an example, consider the causal process of concept triggering as being drawn to 'a zone of stability.' We can base this on Yoshimi's notion of a "stable cognizer" while yet departing from his conception by requiring the apparatus of identity and quantification and thus a symbolic system (2016, p. 18). Once the mind comes close enough through statistical learning, which might comprise a single occasion (consistent with the extremely high rate at which word-learning occurs), Fodorian locking to the property (kind, type, universal etc.) occurs. The asymptotic approximation to a kind at this point activates a symbolic unit of the knowledge system. The knowledge system is based on computational transformations supporting a quantificational syntax and identity semantics. This again must be assumed for the reason that one needs this apparatus to represent kinds as opposed to half-entities corresponding to similarity amalgams; and to stop the infinite regress we need to "come up against kinds" and their corresponding discretely unified meanings (Husserl 2001, p. 244). This can only happen, so far as I know, through a language of thought (Fodor, 1975, 2008). What is statistically learned, therefore, in deep learning, and therefore potentially in our own minds, are representations of experience via "similarity metrics" distinct from a language of thought (Fodor, 2008, p. 158). These contents will be sub-symbolic and serve to trigger the symbolic and discrete units of a language of thought.

Our sketch, as a solution to the arguments, is consistent with the empiricist theory of abstraction and the way in which deep CNNs might learn the representation triangle-*like*. What cannot be learned, however, is:

> [T]he disposition to grasp such and such a concept (i.e. lock to such and such a property) in consequence of having learned such and such a [statistical] stereotype. Experience with things that are asymptotic approximations of The Triangle in Heaven causes locking to triangularity (Fodor, 2008, p. 162).

Locking to triangularity means activation of the identical content conveyed by the concept TRIANGLE. Activation of the concept means the ability to represent a property as holding of all individuals in some domain. It is the ability to represent kinds as such (as opposed to kind-like objects). Such representations are therefore conceptually discrete and symbolically manipulable. We therefore feel forced to combine the sort of learning that deep CNNs are capable of—as an aspect of the similarity semantics of Yoshimian dynamics—with the Fodorian theory of concepts (concept triggering and locking); these together form a hybrid solution to our properly Husserlian problem of how precisely to stop the infinite regress and circularity in content constitution by coming up against kinds. This hybrid two-system solution—a statistical learning system and a conceptual knowledge system—departs from the established empiricism of deep CNNs precisely to the extent that it must attach itself to a "language of thought," which can support "meanings" that are "precise" in the above-required content-identity sense (Pinker, 2007, p. 151).

4 Concluding Discussion: Options for Theorists & Modellers

In this concluding discussion, I want to address the various aims and orientations of modelers and theorists who might want to avoid my solution.

When first presented with the regress/circularity arguments (in section I) there are a few logical options:

1. Ignore the infinite regress/circularity arguments and the logical conditions on content.
2. Accept the infinite regress/circularity arguments but reject the logical conditions on content.
3. Accept the infinite regress/circularity arguments and accept the logical conditions on content.

De facto, the field of neural network theorizing is currently in position 1, because I don't believe the arguments are widely known. They apply to feature-extraction theories, which is what deep CNNs employ. (That is partly why I have focused on deep CNNs.) Choosing 2 could be interpreted as understanding the regress/circularity to be an argument against the existence of logically objective content, given a belief in the default truth of Lockean pooling procedures. In other words, we could say that the success of deep CNNs is proof that Husserlian/Quinean/Fodorian strictures on content constitution are too strong. We can then justify ignoring any and all conditions on content (e.g. the publicity constraint), because we will have demonstrated that excellent recognition is compatible with content only ever being approximated, never identically instantiated (or conceptually conceived *stricto sensu*). We can then work outward toward approximating the logical phenomena of compositionality and systematicity with deep learning systems, as Yoshua Bengio proposes.

There are a couple of reasons, however, why I think we should accept 3. The first reason is that, although the high dimensional representational space of a deep network is opaque to us (e.g. it's not clear what the nodes in the activation patterns of distributed representations are representing) it is nevertheless assumed that the high dimensional spaces *are* representational spaces. If deep learning systems are representation learning systems, and concepts in the logical sense are a kind of representation, then this idea of concepts falls within the explanatory domain of deep learning systems. For example, if it is self-evident that concepts enter into logical relations (as our everyday experience attests) and this entails the properties of compositionality and systematicity (as Husserl insists) these become explananda for any theory of concepts. Yoshua Bengio admits compositionality is an explanandum for deep learning (Bengio, 2019). And Geoffrey Hinton has famously said that deep learning will be able to do everything (Hao, 2020). I'm not aware of anyone denying the phenomena. But one can safely ignore the phenomena by circumscribing the idea of a conceptual representation as a feature-extraction amalgamation. And if one has success with that idea (as with deep CNNs) that does justify ignoring the phenomena to some degree; but I think only for a time.

The second reason is this: the basis of recognition of deep learning is not of individuals as members of a conceptual type. But it seems clear that recognition of individuals as members of a conceptual type is an essential aspect of conceptual meaning. A highly trained, superhumanly recognizing CNN is paradoxically in the position of Quine's pre-individuative child in terms of the possibility of objective representation. Marcus & Davis actually hit on this idea in their polemic against deep learning:

> [Y]ou can train a deep learning system to recognize pictures of Derek Jeter, say, with high accuracy. But that's because the system thinks of 'pictures of Derek Jeter' as *a category of similar pictures*, not because it has any idea of Derek Jeter *as an athlete or an individual human being* (2019, p. 142 italics added).

What I think Marcus & Davis are getting at is the Quinean point, that a deep CNN is a pre-individuative (and therefore pre-conceptual) machine. Note the reference to similarity semantics. A deep CNN is trapped in the Quinean pre-individuative stage to the degree that its knowledge (its competence) is based on similarity semantics (Firestone, 2020). It cannot escape similarity semantics due to the regress/circularity in content constitution.

Assume for the moment that this line of argument is essentially correct—that statistical learning must be supplemented by a LOT to avoid the regress/circularity in content constitution. Then, provided the aims are different, the nature of deep CNNs, and what they reveal about statistical learning, can potentially be incorporated into a general theory of mind—perhaps explaining prototypicality judgements based in similarity. I have nothing against this. All I am saying is that the logical nature of concepts and object identity are correlative explananda, along with their entailments (e.g. compositionality), for representation theory. Since deep learning is a highly successful branch of representation theory, it becomes a question whether these phenomena are proper aims of the theory behind deep CNNs, which employ the empiricist theory of abstraction. And I am saying that if deep CNNs really are the mechanical realization of the empiricist theory of abstraction, then these phenomena will be missed due to the infinite regress/circularity, and are therefore not proper aims of the theory behind deep CNNs.

It's worth, therefore, discussing what the aims of modelers and theorists are in deep learning. It may be that these aims are very modest. Perhaps concepts and object identity and the associated logical phenomenology characteristic of thought (systematicity, compositionality) are not explananda for them. Perhaps they use this language loosely, without intending to mean what these words mean in philosophical discussion (cf. Machery, 2009). If so, they can ignore my arguments for supplementation in terms of the familiar apparatus of a language of thought. I wish all such modelers and theorists the best of luck—I am not against loose-talk per se.

If, however, these words are to be taken seriously as theoretically subsuming all conceptual phenomena, then I believe this hinders progress toward a general theory of mind. And that's only because the use of the terms 'concept' and 'object identity' in discussions of deep learning systems obscures the representational phenomena at issue that are relevant to a general theory of mind/intelligence (e.g. systematicity, compositionality, etc.); and at least some deep learning modelers (e.g. Bengio) admit these are real explananda that must be faced eventually.

One distinction that might clarify the theoretical gulf between the system's aims and the relevant explananda is the distinction between symbolic and sub-symbolic content. Shea (2021) has recently argued that deep learning systems are probably a good model for the transition from sub-symbolic contents (textures, colors etc.) to the tokening of concepts (e.g. CAT). This registers the direction of discussion since Fodor's *Concepts* (1998): "If, looking at Greycat, I take him to be a cat, then too I apply the concept CAT to Greycat" (24). I contrast Shea and Fodor here to bring out how the current discussion assumes that concepts are feature-extraction amalgamations based on sub-symbolic contents. The infinite regress/circularity arguments apply to feature-extraction amalgamations. But again this is only a problem if symbolic content (a logical conception of CAT) capable of traditional computational operations is a goal; it may not be.

Part of my argument has been that a logical concept of, say, CAT should be a goal also for object-side reasons. A sub-symbolic system represents quasi-objects based on segmentation of contours or patterns in the input, the sensory presence of bodies, memory and recognition of the sensory presence of bodies, etc. The subjectivity of similarity-based objects arises for objects that are *essentially relative* (not just accidentally) to the past history of the individual organism(/machine) (see Quine, 2004, p. 290). Deep learning proposes to overcome this subjectivity by brute force, with extremely large data-sets. Nevertheless, strict object identity will always require quantification and discrete symbolic units, if the infinite regress/circularity arguments hold. Lacking these representational resources, a representational being/machine can only be said to represent a sort of half-entity corresponding to a similarity amalgam. This is not a metaphor—it's a technical term (to be taken quite literally), meaning a subjective representation, incapable of logic and reasoning, because based on the amalgamated history of semantic segmentation or representation of similarity 'clumps' in the environment (Burge, 2010, p. 236; cf. Millikan, 2017). It is sub-symbolic (Smolensky, 1991). This is important for our argument since Quinean subjective representations correspond to the representations of deep CNNs, with the difference that deep CNNs have far more data at their disposal. I merely extend Quine's argument and group deep

CNNs with his notion of animals and pre-linguistic infants *despite* deep learning's recognitional prowess far exceeding the limits of all known organic creatures (cf. Cangelosi & Schlesinger, 2015). But I further argue that there is no bridge from this kind of representation to the kind 'required' for concepts and object identities due to the arguments. I conclude that the only solution given the arguments is to suppose a language of thought (see Sect. III). If one is not interested in the phenomena (the logical phenomenology) related to concepts and object identities in the strict sense, one can ignore the arguments (per the above option).

We can summarize the broad outlines of this concluding discussion with the following:

1. Deep learning systems, in particular deep CNNs, employ the empiricist theory of abstraction (Buckner, 2018).
2. The empiricist theory of abstraction generates content on the basis of induction by noting similarities and differences among phenomenal features of objects.
3. These similarities and differences are defined, in deep learning, over pixels, sound images etc. (sub-symbolic contents).
4. Deep learning systems, in particular deep CNNs, therefore induct sub-symbolic similarity amalgams.
5. Sub-symbolic similarity amalgams are essentially contrasted with symbolic contents susceptible of an apparatus of identity and quantification.
6. Therefore, insofar as identity semantics (concepts and their objects qua identity) are relevant explananda for deep learning (as is admitted by all sides), a supplementary mechanism involving the apparatus of identity and quantification is required (Marcus & Davis, 2019).

Premise 5 is given support by the infinite regress/circularity arguments (part I). The need for the apparatus of identity and quantification is supported by Fodor's much discussed 'publicity constraint' on conceptual content (Prinz, 2002; Edwards, 2009; Schneider, 2011). As to Premise 6, it should be noted that this is textually true—"concept… and object identity" are explicitly considered *explananda* of deep learning (Goodfellow et al., 2016, p. 17; Kelleher, 2019). I do not think that these authors have reflected that these terms are connected with systematicity, compositionality, and other recognized cognitive explananda. I argued in part I that they are. And since deep learning is tied to the empiricist theory of abstraction (Buckner, 2018), which can neither result in concept nor object identity due to the resultant similarity semantics (Fodor, 1998, Husserl, 2001), deep learning, as applied to the mind, will, again, need to be supplemented in terms of the computational-linguistic apparatus normally associated with a language of thought.

The duty of the philosopher, I think, is to clear a path for the scientist to discover a mechanism. One can ignore the phenomena and the arguments, but if one takes them seriously, one will necessarily seek the discovery of mechanisms that support "transformations needed for quantification," potentially yielding an identity semantics capable of representing kinds (types, properties, etc.) (Hinzen, 2006, p. 177). Such mechanisms have been sought with some success (O' Reilly et al., 2014); and the language of thought may even have a readily interpretable neuroscientific realization (Gallistel, 2018). Our solution to the circularity and infinite regress posed by the similarity semantics of deep learning is to suppose that it must be supplemented by a separate system defined by specific, unified meanings, detachable in principle from statistically induced degrees of associative fulfilment (Wieskopf, 2015). A place for deep CNNs is included in our solution, to the degree that such functions for similarity learning are a real aspect of statistical learning in experience, as well as the informational basis of similarity judgments. In future work, I hope to develop this unified picture of the phenomenology of logical experiences—the original sense of (Husserlian) phenomenology—with the language of thought and deep CNNs as a full explanatory theory consistent with the above arguments.

Acknowledgements This article was originally an Eastern APA Symposium paper delivered on January 6th, 2022. Eli Alshanetsky, Andrew Tyler Richmond, & Christian Carrozzo contributed criticisms and ideas that brought the article to its final form.

References

Bengio, Y. (2019). *Towards compositional understanding of the world by deep learning*. Peking University.

Buckner, C. (2018). Empiricism without Magic: Transformational abstraction in deep convolutional neural networks. *Synthese*, *195*, 5339–5372.

Burge, T. (2010). *Origins of objectivity*. Oxford.

Cain, M. J. (2016). *The philosophy of cognitive science*. Polity.

Cangelosi, A., & Schlesinger, M. (2015). *Developmental robotics: From babies to robots.* MIT Press.

Carey, S. (2009). *The origin of concepts*. Oxford.

Churchland, P. (2012). *Plato's camera: How the physical brain captures a Landscape of Abstract Universals*. MIT Press.

Dube, S. (2021). *An intuitive exploration of artificial intelligence: Theory & applications of deep learning*. Springer.

Edwards, K. (2009). What concepts do. *Synthese, 170*, 289–310.

Firestone, C. (2020). Performance vs. competence. *Human-Machine Comparisons*. https://doi.org/10.1073/pnas.1905334117

Fodor, J. (1975). *The language of thought*. Harvard.

Fodor, J. (1998). *Concepts*. Oxford.

Fodor, J. (2008). *LOT 2: The language of thought revisited*. Oxford.

Fodor, J., & Pylyshyn, Z. (2015). *Minds without meanings.* MIT.

Fodor, J., Ernest Lepore. (1992). *Holism: A shopper's guide*. Blackwell.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics, 36*, 193–202.

Gallistel, R. C. (2018). The neurobiological bases of the computational theory of mind. *On concepts, modules & language.* Oxford University Press.

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

Hao, K. (2020). *AI pioneer Geoff Hinton: Deep learning is going to be able to do everything*. MIT Tech Review.

Hinzen, W. (2006). *Mind design & minimal syntax*. Oxford.

Hopp, W. (2011). *Perception and knowledge: A phenomenological account*. Cambridge.

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology, 28*, 229–289.

Husserl, E. (2001). *Logical investigations*. Routledge.

Kelleher, J. (2019). *Deep learning*. MIT Press.

Krizhevsky, Alex, & I., Sutskever, & Hinton, G. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems, 25*, 1.

Machery, E. (2009). *Doing without concepts*. Oxford.

Marcus, G. (2001). *The algebraic mind*. MIT.

Marcus, G., & Davis, E. (2019). *Rebooting AI.* Pantheon.

Millikan, R. (2017). *Beyond concepts*. Oxford.

O'Reilly, R. C., Petrov, A., Cohen, J., Lebiere, C., Herd, S., & Kriete, T. (2014). How limited systematicity emerges. *The architecture of cognition: Rethinking fodor & Pylyshyn's systematicity challenge* (pp. 191–225). MIT.

Pinker, S. (2007). *The stuff of thought*. Viking.

Prinz, J. (2002). *Furnishing the mind*. MIT.

Quine, W. V. (2004). *O. Quintessence*. Belknap.

Schneider, S. (2011). *The language of thought*. MIT.

Sejnowski, T. (2018). *The deep learning revolution*. MIT Press.

Shea, N. (2021). Moving beyond content-specific computation in artificial neural networks. *Mind & Language*. https://doi.org/10.1111/mila.12387

Simons, P. (1995). Meaning and language. *The Cambridge companion to Husserl* (pp. 106–137). Cambridge.

Smolensky, P. (1991). Connectionism, constituency and the language of thought. *Minds, brains, and computers* (pp. 286–306). Blackwell.

Smolensky, P., & Legendre, G. (2006). *The harmonic mind.* MIT.

Strogatz, S. (2015). *Nonlinear dynamics and chaos*. Westview.

Weiskopf, D. A. (2015). Observational concepts. *The conceptual mind* (pp. 223–248). MIT.

Yoshimi, J. (2011). Phenomenology & connectionism. *Frontiers in Psychology* (2011). https://doi.org/10.3389/fpsyg.2011.00288

Yoshimi, J. (2016). *Husserlian phenomenology: A unifying account*. Springer.

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