Mad-Dog Nativism, Neural Networks, and The New Riddle of Concept Learning [1]

Sam Clarke
Departments of Philosophy and Psychology
University of Southern California
sam.clarke@usc.edu

Draft as of Dec 29th, 2025 – comments welcome.

This paper contrasts two arguments for concept nativism. The first is Fodor's infamous argument for "mad-dog nativism", the view that all or virtually all concepts are innate. I argue that recent work in computer science undermines Fodor's argument, but not because it reveals Fodor's conclusion to be absurd or untenable (as many commentators assume). The real reason why Fodor's argument is ineffective is that – while its conclusion reflects a serious empirical possibility – it motivates a view which is nativist in name only, allowing concepts to be acquired in ways which many contemporary empiricists would welcome. Against this backdrop, I proceed to develop a lesser known, second argument for concept nativism: *The New Riddle of Concept Learning*. I show that this argument avoids the shortcomings of Fodor's mad-dog nativism, dictating a more robust brand of nativism in myriad contested conceptual domains – including that of natural number and theory of mind – without spiraling into absurdity and implying that concepts like CARBURETOR are similarly innate.

0. Introduction

Concepts are the building blocks of thought; the sub-sentential constituents from which complex ideas are constructed. On mainstream cognitivist accounts, this is because thoughts are, quite literally, composed from concepts. The meaning of a truth-evaluable thought, like KEVIN LOVES DOGS, is fixed by the meaning of the concepts it comprises (KEVIN, LOVES, DOGS) plus the mode by which these are combined (Fodor 2008). Since the concepts we possess thereby delimit the thoughts we can entertain, the question of *how* we acquire concepts is among the most pressing in philosophy of mind. But, where an old and venerable rationalist tradition has argued that many concepts must be innately bestowed upon us by God or biology, recent advances in artificial intelligence and cognitive development have been taken to undermine these suggestions. For instance, neural networks, which appear to abstract novel representations from their training data, have been taken to vindicate the contrasting empiricist conviction that experience can and does expand our conceptual horizons without innate structuring of the proposed sort.

The present treatment considers reasons to resist this recent wave of anti-nativist enthusiasm. In Sections 1 and 2, I argue that advances in artificial intelligence *do* undermine one prominent motivation for concept nativism – namely, Fodor's argument for "mad-dog nativism", according to which all or virtually all concepts must be innate. This is not because Fodor's conclusion is absurd, as many commentators assume – I think it reflects a serious empirical possibility. The problem is, rather, that recent artificial intelligence demonstrates that Fodor's conclusion would merely constitute a pyrrhic victory for the nativist, allowing concepts to be acquired in ways which most contemporary

[1] For helpful (often detailed) comments and discussion, I would like to thank Luca Barlassina, Zoe Drayson, Gabe Dupre, Veronica Gómez Sánchez, Steven Gross, Kevin Lande, Henry Schiller, Alexis Wellwood, and an audience at UC Davis.

empiricists/anti-nativists would welcome. In response, Section 3 advances an improved argument for concept nativism that avoids these concerns. This improved argument – *The New Riddle of Concept Learning* – recommends a robust brand of concept nativism in contested domains, like that of natural number and theory of mind, without spiraling into absurdity and implying that *all* human concepts (e.g., CARBURETOR) are similarly innate. The result is a powerful yet underappreciated motivation for a strong yet plausible brand of concept nativism; one which rebuts non-nativist models of concept acquisition inspired by recent artificial intelligence and developmental science.

1. Fodor's Riddle of Concept Learning

Fodor's 'mad-dog nativism' – the view that virtually all human concepts are unlearnt and innate – took varied forms throughout his career (see Laurence & Margolis [2024: Ch.24] for an overview). But who cares? Fodor's proposal strikes many commentators as sufficiently absurd as to be a reductio of itself. Why would anyone entertain the bizarre claim that concepts, like PAINT, SPATULA, and T-REX, are innate in the sense of being unlearnt, let alone spend a career defending it?

Despite oft-claimed implausibility, Fodor's arguments for mad-dog nativism were lucid and instructive. Each began from a straightforward observation: there are ways in which a concept could be *acquired* which would not involve it being *learnt*. To illustrate, suppose that Lucky Luke lacks a concept C at a time *t1* but later acquires C through a chance blow to the head that just so happens to restructure his brain appropriately. Luke would have thereby acquired a new concept. But intuitively, he wouldn't have acquired it through *learning*. Alternatively, we can imagine a nearby future in which neurosurgeons rewire a chimpanzee's brain, such that it comes to possess new concepts it previously lacked (e.g., HIGGS-BOSON). This might enable Super-Chimp to do surprisingly well on a physics exam. But, as with the previous example, it feels odd to describe this as a case in which Super-Chimp has *learnt* any new concepts – intuitively, they merely *acquired* these. Indeed, we could consider the concept acquisition process an active occurrence, occurring at the psychological level, without taking it to involve learning. For if novel concepts of, say, fauna and car parts were to pop into Sid's head whenever he actively reflects on the pain in his stubbed toe, we might think of his actively attended pain state as somehow causing these concepts' acquisition. But, even thought of in this way, it remains highly irregular to think of these concepts as being *learnt*. This is analogous to the fact that if backpain causes me to form particularly pessimistic beliefs about the political situation, my pain might be said to have 'triggered' these new beliefs without implying that they were learnt on this basis. Of course, no one thinks that human concepts are routinely acquired in any of these ways. Such fanciful examples simply highlight an intuitive distinction in need of explication, prompting the question: What distinguishes concept *learning* from concept *acquisition* more generally?

Fodor's answer to this question was, again, straightforward. He proposed that for a concept to be *learnt*, it would need to be acquired through a *rational process of evidence-based hypothesis testing*. To motivate this claim, note that this is lacking in the above examples – cases which, intuitively, failed to involve learning. Take Lucky Luke. In acquiring a new concept C, Luke acquired an ability to think thoughts comprising C. However, the blow to Luke's head, that was responsible for his acquisition of C, failed

to provide Luke – the subject – with any evidence that C is about *C* as opposed to *C*,* or something else entirely. For instance, if C is the concept PAPAYA, Luke's head trauma will have provided Luke – the person – with no reason to hypothesize that C is about papaya, as opposed to the broader genus of carica plants, some unrelated category, or nothing at all. In this way, C's acquisition did not involve Luke rationally formulating or evaluating hypotheses about C's meaning in light of evidence at his disposal. Ditto for the concept acquisition processes described with respect to Super Chimp and Sid.

Fodor's more tendentious claim was that this makes these examples quite unlike paradigm cases in which something is *learnt*. To see why one might say this, forget about concepts for a moment; imagine that you are a child, *learning* the meaning of the word "fish". This might begin when your parents take you to an aquarium, point at certain objects, and say things like "Fish!", "Look at the pretty fish" or "That fish would make a tasty dinner!". In each case, you can ask yourself *what are these 'fish' that Mum and Dad keep referencing?* prompting you to project and rationally evaluate certain hypotheses. For instance, you might note that each time "fish" is uttered, the object being pointed at is an aquatic creature and thereby hypothesize that "fish" means AQUATIC CREATURE. Of course, this hypothesis is false, since its application leads to the misclassification of squid and whales as fish. But while false, the hypothesis isn't outlandish: it reflects a rational response to evidence at your disposal. And if its shortcomings are pointed out, you could respond by reformulating your hypothesis. For Fodor, this is because *learning* the word's meaning, as opposed to merely *acquiring* some appreciation for this, involves you engaging in a rational or reasoned process of hypothesis testing wherein hypotheses as to the word's meaning are projected and subsequently tested in response to evidence.

It's worth stressing that Fodor sometimes suggested that this much is characteristic of learning, writ large, be it the learning of novel word meanings or dance routines.[2] He also routinely asserted that it is just plain difficult to see how concept learning would proceed if not through some form of reasoned projection and hypothesis confirmation; as Fodor saw it, this is "the only kind of theory that has ever been proposed for concept learning" (Fodor 1975: 36). But irrespective of whether Fodor was right to say any of this (complications forthcoming), he was correct to observe that this much is often presupposed by proponents of concept learning. For instance, Fodor stressed that when psychological behaviorists sought to study concept learning, they employed experimental paradigms in which participants were given made-up words and tested for their ability to correctly identify their (stipulated) meanings in response to examples that either do or do not fall under the corresponding concept's extension. Thus, participants might be given a made-up word, "flurg", and told that a brown square is flurg and that a red triangle is not flurg, etc., such that they can begin formulating hypotheses about the word's meaning – responding to this evidence and hypothesizing that 'flurg' might mean DULL COLORED SHAPE*,* or BROWN SQUARE*,* etc. The researchers would then study how these hypotheses are revised and updated in light of further evidence – e.g., how participants respond to then being told that a brown pentagon is flurg. This might continue until the learner arrives at the

[2] Laurence and Margolis (2024) object that *rote learning* need not involve hypothesis testing. If true, Fodor's (still contentious) claim, is best understood as *learning of a sort that plausibly underlies concept learning relies on hypothesis testing.* But see Antony (2020) for helpful discussion of this point.

intended interpretation of the word, operationalized as correct categorization on >N% of trials or X successive trials, at which point *voila*: they are said to have *learnt* the meaning of the word and its corresponding concept, FLURG. Indeed, while this is presented as typical of concept learning studies run by behaviorists of old, contemporary cognitivist accounts frequently fit this mold. For instance, in a much-cited paper, that is representative of the "rational constructivist" program, popular with Bay Area developmentalists, Perfors et al. (2008) present a "tutorial introduction to Bayesian models of cognitive development" in which conceptual development proceeds via the explicit projection and Bayesian updating of hypotheses concerning the meaning of novel concepts children acquire.

The trouble is: Once this is conceded, Fodor objects that the very idea that new concepts might be learnt now seems close to incoherent. If we accept that concept learning is not identical to concept acquisition, and we accept that part of the difference resides in the fact that concept learning proceeds via the explicit projection and testing of hypotheses as to the to-be-acquired concepts' meanings, then the concept learner had better be able to formulate the relevant hypotheses to even initiate the learning process. For instance, if learning the meaning of CHAIR involves landing on the hypothesis that CHAIR = *four-legged object apt to be sat on* (c.f., Elbourne 2011), then the learner had better have the competence to represent *four-legged object apt to be sat on* to themselves prior to engagement in the learning process. But, if learning a novel concept requires that one already has the competence to represent its meaning to oneself, this sounds tantamount to saying that one must – effectively – already possess the concept whose acquisition we wished to explain. Hence, *learning* a new concept requires that we already possess that very concept!

We can distinguish two versions of Fodor's argument. A 'modest' reading, recommended by Fodor (1975; 1981), holds that once the abovementioned logic is appreciated, it follows that one cannot learn new concepts *which expand the expressive potential of one's conceptual system*. Suppose that Conceiving Connie has at her disposal a finite stock of just 100 primitive concepts, C^1-C^{100}. Experience might teach her to put any number of these pre-existing concepts together in novel ways – thus, she might "learn" through a process of hypothesis testing that the meaning of PAINT is identical to the meaning of her pre-existing concepts, C^4 and C^9 (such that something is PAINT just in case it is $C^4 + C^9$). What it would not enable her to do is formulate hypotheses as to the meaning of a concept that cannot be expressed in terms of C^1-C^{100}. Hence, if a concept C^{101} cannot be reduced to any combination of C^1-C^{100}, then Connie simply will not be able to entertain it, no matter how hard she tries. Her learning of C^{101}, through a process of hypothesis testing, will be strictly impossible.

At first blush, this might not sound so bad. It is, for instance, consistent with the ambitions of an Enlightenment empiricist, who seeks to reduce all human concepts to associated combinations of innate sensory primitives. But aside from the fact that it's notoriously difficult to see how one might reduce abstract concepts, like KNOWLEDGE or JUSTICE, to associated arrangements of sensory primitives – e.g., retinotopically arranged blobs of phenomenal color – it's hard to even see how superficially observable properties, like TRIANGULARITY or GREEN, could be so reduced. For if one only ever experiences specific triangles or specific shades of green, each with its own idiosyncratic

dimensions and shading properties, it's unclear how this would provide the resources to abstract away from these idiosyncrasies and arrive at a concept TRIANGLE or GREEN that represents triangles or green in full generality. Thus, Fodor's arguments could already be taken to show that the range of innate conceptual primitives needs to be much richer than traditional empiricism allows. But really this is just the start. If one shares Fodor's general skepticism about the definability of *any* monomorphemic lexical concepts, or the identification of concepts with data-structures more broadly (e.g., prototypes), then the range of innate and unlearnt conceptual primitives will have to be vast – it will have to include all the monomorphemic lexical concepts humans acquire, including PAINT, CHAIR, and CARBURETOR (Fodor 1975; 1981). In fact, it will have to include all the monomorphemic lexical concepts that humans ever *could* acquire.

And yet, this lofty conclusion must still be distinguished from an even stronger reading of Fodor's argument. On this second reading, the (un)definability of lexical concepts is a moot point. Notice that while the preceding argument recommends that *primitive* concepts cannot be learnt, it still allows for the learning of *complex* concepts. For instance, if Connie has among her concepts C^1-C^{100} BROWN and COW, she might formulate a hypothesis as to the extension of a new complex concept BROWN COW that is identical to their conjunction (hypothesis: BROWN COW = BROWN + COW). Given this, one might suppose that the crux of Fodor's argument is simply to find a way of rejecting his conceptual atomism, and to find some way of defining the concepts one thinks must be learnt (e.g., CARBURETOR) in terms of a palatable base of innate primitives.

By 2008, Fodor was arguing that this misses the point. Suppose again that Connie has the concepts BROWN and COW but has never formulated the complex concept BROWN COW prior to seeing Maisy. That Maisy is a brown cow causes Connie to token a previously un-tokened complex concept, BROWN COW, combining her pre-existing primitive concepts BROWN and COW in previously unthought ways. What Fodor now submits is that this still doesn't involve Connie *learning* this complex concept. Why not? Because the initial tokening of the complex concept does not, itself, emerge via rational hypothesis testing. Seeing Maisy might trigger activation of a previously unformulated complex concept. But to formulate a hypothesis that BROWN COW is identical to BROWN + COW, BROWN + COW must already be accessible to Connie's hypothesis-generating mechanisms. In other words, the initial formulation of BROWN COW cannot emerge through hypothesis testing since its formulation is a prerequisite for use in hypothesis testing. Hence, where the first reading of Fodor's argument recommends that all monomorphemic lexical concepts must be innate in the sense of being unlearnt (Fodor 1975; 1981), the second reading maintains that even manifestly complex concepts, like PINK OR GREEN FAIRY CAKES, must be innate in much the same way (Fodor 2008).

2. The Challenge from Neural Networks

You don't need to be John Locke to find Fodor's advertised conclusions uncomfortable. His argument, in either its weaker or stronger incarnation, purports to show that all or virtually all concepts must be innate in the sense of being unlearnt, even those which pertain to recent technologies (e.g. WALKIE-TALKIE), posits (e.g., QUARK) and discoveries (e.g., SYNGNATHIDAE) – i.e.,

concepts which presumably never became active in the minds of any humans for the first 299,900+ years of our estimated 300,000-year existence. If this implied that all such concepts, categories and kinds were anticipated by genetic evolution, back in the Pleistocene, as Putnam (1988), Sterelny (1989), and many others suggest, I agree that this would be absurd.

The thing is: Fodor's arguments show nothing of the sort. Suppose we bracket attempts to address Fodor's arguments head on (Antony 2020; Beck 2017; Carey 2009; Laurence & Margolis 2002; Weiskopf 2008) and accept – if only for the sake of argument – their logic, concluding with Fodor that novel concepts really cannot be learnt. What would this show? It would show that concepts are acquired in some other way. This might involve certain concepts' hardwiring via genetic evolution. But as the examples of Lucky Luke, Super-Chimp, and Sid illustrate, it might also involve novel concepts being acquired throughout the lifespan. In fact, it might even involve novel concepts being acquired throughout the lifespan in ways that are flexibly shaped by one's experience, without it following that their acquisition involves 'learning' as Fodor understands this.

I'll illustrate this point using an example from computer science. The example might seem to straightforwardly refute Fodor's arguments. However, I will show that Fodor has the resources to dismiss the concern. An upshot is that *mad-dog nativism* – understood as Fodor's claim that *all* concepts are unlearnt – is not as far-fetched as most commentators assume. It is a coherent scientific hypothesis that is simply at the mercy of empirical fortune, at least if we accept Fodor's analysis of what concept *learning* requires. But while this sounds like good news for mad-dog nativism, I'll proceed to note a more fundamental problem with the view: Namely, that it is nativist in name only, allowing novel concepts to be acquired in ways that most contemporary empiricists should celebrate. Thus, I will argue that while mad-dog nativism reflects a serious possibility regarding the origin of concepts, it talks past its opponents. The result is that a concept nativism that's worth its salt will need to be motivated quite differently. Some such motivation is what I'll aim to provide in Section 3.

2.1 A Case Study from Computer Science

In a celebrated *Science* paper, Nasr et al. (2019) describe a convolutional neural network that was trained on 1.2 million labelled images from the ImageNet dataset. These images corresponded to 1,000 pre-labelled categories (e.g., *dog, necklace,* etc.). After training, the network was tested for how well it would label 50,000 new images, not found in the original dataset, into the 1,000 categories it had been trained on. The network labelled just 49.9% of these accurately – not quite the super-intelligent singularity some fear to be nigh, but statistically significant levels of performance regardless ($P<0.001$).

More interestingly for us, Nasr et al.'s network proved to be *more than a mediocre image categorizer*. Having merely been trained for image classification in the abovementioned ways, Nasr et al. proceeded to test the network on a number-estimation task. This was motivated by their conjecture that object categorization is facilitated by numerical representations in the mammalian visual cortex (p.1). To test this, the researchers presented the trained network with images containing 1-30 dots and, without further training, found that neurons in the network's feature-extraction network were already tuned

to approximate numbers or 'numerosities'. Thus, after the network had been trained to categorize images in its training data, individual neurons in the network were found to reliably fire in response to collections containing approximately seven dots, or 22 dots, or 13 dots, etc. This was observed even though non-numerical properties of the displays (e.g., the cumulative and/or average area of the dots) were controlled for (but see Odic MS. for concerns). On various theories of content ascription – including that of Fodor (1994) – this could be taken to indicate that individual neurons in the network were representing approximate numbers/numerical contents (in the sense defended by, e.g., Clarke & Beck 2021). But, if this is granted, it could seem to demonstrate that atomic representations/concepts really can be learnt, *pace* Fodor. This is because the network (as a whole) did not start out representing numbers, approximate or otherwise. Its innate representational primitives corresponded to pixel values of a sort encoded in its input images plus 1,000 non-numerical object categories.[3] Thus, numerical representations were seemingly acquired by the network, through a process of *deep learning*. For many contemporary empiricists, I suspect that this would serve as an existence-proof that learning from experience can expand the expressive potential of one's conceptual system, thereby refuting both stronger and weaker readings of Fodor's argument.

But is this really a refutation of Fodor's mad-dog nativism? The devil's in the details.

On the one hand, it would be disingenuous to deny that results of the abovementioned sort constitute a boon for empiricism. Such results suggest that novel representations can be acquired from something akin to experience. And while it is true that classical empiricists suggested that abstract concepts were defined out of primitive sensory constituents, and hence reducible to associated combinations of these, it was at least as common to find them proposing that concepts are learnt through processes of "abstraction" (Berkeley 1710/1975; Hume 1739/1978; Reid 1785/2002; Mill 1882; James 1890; Russell 1912). Roughly: Processes in which general concepts (e.g., RED) are acquired when the subject observes determinate examples (e.g., RED^{17} and RED^{52}) and, somehow, appreciates the relevant similarities between them (e.g., that all are instances of RED). Indeed, the arch-empiricist, John Locke maintained that this is how "all General *Ideas* are made" (1690/1975). But while rationalist commentators dismiss this as an 'illusory explanation' of concept acquisition (Laurence & Margolis 2024: Ch5), owing to the obscurity of thinking that one could notice *the* relevant similarity among varied instances of a category (e.g., RED^{17}; RED^{52}; etc.) prior to possessing a general concept that each falls under (e.g., RED), it is plausible to suppose that this is what Nasr et al.'s network achieved (see: Buckner 2018). After all, it is hard to see how representations of numerical quantity, like SEVEN or SEVEN-ISH, could have been composed out of the network's pre-existing

[3] A complication: Nasr et al.'s network was implemented on a digital computer, which represented numbers prior to training, using these to express weightings between neurons and neurons' activation functions. Nevertheless, these numerical representations did not feature in a hypothesis space on which number-tuned neurons' numerical contents were framed – rather, backpropagation, driven by stored object knowledge, resulted in neurons reliably responding to approximate numerical quantities. Moreover, while the network was implemented on a digital computer, this was merely a convenience. Nasr et al. could have constructed the same network from a vast array of interconnected physical nodes in a warehouse. In that case, hyperparameters, like nodes' activation functions, and their weightings would not be represented by numbers, though numbers could be used to describe their properties. Nevertheless, the response profile of the neurons would be identical post-training, and number-sensitive neurons would emerge in much the same way.

representations of object categories and pixel values (such that SEVEN or SEVENISH = NECKLACE & BLUE[16] &… say). Rather, the network's training simply led to its descending upon an object-recognition function that happened to employ numerical representations because this improved performance at the output layer. This, plausibly, resulted from it abstracting numerical contents following exposure to varied images in which the target objects had the associated numerical properties (e.g., *four* or *fourish* legs; *four* or *fourish* sides; etc.). So, despite Fodor's (1975; 1981) arguments, neural networks like Nasr et al.'s indicate that conceptualization need not be limited to the expressive potential of one's initial representational repertoire: Neural networks plausibly abstract previously unrepresentable categories from their training data in something like the way empiricists maintain. Thus, it's no wonder that connectionist architectures have often been advocated by those with empiricist leanings (Clark 1993; Churchland 1986; Prinz 2013) and leapt upon by card-carrying empiricists as plausible accounts of concept learning (Buckner 2023; Clatterbuck & Gentry 2025).

On the other hand, Fodor would respond that this is entirely beside the point. For, even if neural networks like Nasr et al.'s acquire new representations/concepts, Fodor spent his career defending positions which would lead him to deny that they do so through *learning*.

For a start, Fodor argued that concept *learning* requires that learners make reasoned or rational responses to evidence at their disposal. This is contentious but builds upon the less controversial insight that learning is, ultimately, a psychological process (Samuels 2002; Shea 2010; Weiskopf 2008). As such, it is notable that Fodor was at pains to argue that non-classical neural networks, of the sort employed by Nasr et al., provide implausible models of human psychology. For instance, Fodor and Pylyshyn (1988) argued that insofar as non-classical neural networks provide insight into the workings of the human mind, it is by providing an account of how human brains implement a radically different classical architecture at an autonomous level of psychological description. So, even if networks, like Nasr et al.'s, provide a viable account of how new concepts/representations can be abstracted from experience, Fodor would maintain that this abstraction is occurring at a level of description that is wholly irrelevant to that at which learning would need to obtain. Indeed, Fodor (2008) came close to anticipating this point: While he continued to view his arguments against concept learning as decisive, he emphasized that this is consistent with the acquisition of novel concepts occurring throughout the lifespan, at sub-psychological levels of neural implementation.

This is to say nothing of the fact that, even if Fodor's views on the inadequacies of neural networks are rejected, it is hard to see how neural networks like Nasr et al.'s could be engaged in rational hypothesis testing when acquiring new representations – i.e., engaged in *learning* as Fodor understands this. Insofar as Nasr et al.'s system engaged in evidence-based projection and updating of hypotheses, it is natural to think that it did so with respect to the meanings of the pre-represented object categories it was trained to map input images onto. Recall that during its training phase, Nasr et al.'s network was presented with images, which it categorized as NECKLACE, DOG, RAMEN, etc. according to 1,000 categories its training algorithm already encoded. A backpropagation algorithm then flagged these categorizations as correct or incorrect, drawing on pre-programmed 'knowledge' of correct

answers, enabling it to alter the weightings among neurons in the network (strengthening connections that resulted in correct categorizations and weakening connections which led to misclassifications). Squinting only slightly, these connections might be seen to embody hypotheses about the appropriate interpretation of its inputs, with this process of backpropagation rationally updating these. But even granting this much, evidence of the form "Yup, that's a pigeon", "Nope, not a hotdog", simply does not seem to confirm, refute, or bear upon hypotheses as to the meaning of the numerical concepts/representations the system acquired (e.g., SEVEN or SEVENISH). Hence, it remains difficult to see how the network could have *learnt* to produce these novel representations, as Fodor argues "learning" must be understood.

2.2 Why Nativists Need a New Argument

There's more to be said on these matters. Evaluating these mad-dog nativist-friendly-responses to cases of neural network-based abstraction would be a Herculean task. It would require consideration of the grounds for thinking psychology an autonomous discipline (Fodor 1974; 1997), the grounds for thinking connectionist networks inadequate as models of cognition (Fodor & Pylyshyn 1988; Fodor & McLaughlin 1990), and a deep-dive into the learning algorithms that neural networks employ (Kelleher 2019, ch.5-6); not to mention careful analysis of what hypotheses, rationality and epistemic justification amount to (Antony 2020). But, without delving into these complexities, we are poised to zoom out and appreciate two general observations on the state of the dialectic.

The first is that mad-dog nativism, understood as the claim that novel concepts can't be *learnt* (in Fodor's technical sense of the term), isn't as outrageous as it's made out to be. Rather than being straightforwardly absurd (Churchland 1986; Putnam 1988; Sterelny 1989), it's plausible that biologically inspired neural networks show how new concepts might be flexibly acquired through the lifespan, in ways that are consistent with them being unlearnt, at least as Fodor understands matters. To be clear, this is no foregone conclusion. My point is simply that, even if there are problems with Fodor's argument, and it transpires that concept learning is not straightforwardly confused in the ways Fodor suggests, additional work is needed to show that novel concepts are – in fact – acquired through processes apt to be called learning.

More importantly, however, nativists would be wrong to pop the champagne. Even if the above points hold, and mad-dog nativism emerges as a sensible hypothesis, the worry now looms that Fodor has motivated a view that is nativist in name only. Many card-carrying empiricists happily avail themselves of advances in neural network architectures when motivating theories of cognitive development (Buckner 2023; Churchland 1986; Clark 1993; Clatterbuck & Gentry 2025), even taking these insights to vindicate the Enlightenment empiricist's notion that novel concepts are learnt via abstraction (Buckner 2018). This is to say nothing of the fact that modern empiricists typically reject Fodor's claims regarding the autonomy of psychology from neuroscience and his analysis of 'learning' – i.e., claims that insulate Fodor from the challenges modern neural networks pose. In this way, a contemporary empiricist would likely reject Fodor's characterization of "learning" or "innate" as irrelevant in this context. In any case, once it's conceded that novel concepts can be flexibly acquired

through experience, via sub-psychological processes, that arguably look like empiricist abstraction, any mad-dog nativism which survives looks surprisingly pedestrian. It allows concepts to be acquired in ways that would make most contemporary empiricists quite happy.

3. The New Riddle of Concept Learning

Fodor's arguments for concept nativism are ineffective. This is not because mad-dog nativism is too absurd to take seriously (*pace* Churchland 1986; Putnam 1988; Sterelny 1989). Nor is it that intelligible theories of concept learning might be formulated (Beck 2017; Carey 2009; Laurence & Margolis 2024; Weiskopf 2008), as if this automatically implies that these should be preferred to Fodor's hypothesis. The fundamental problem is that, even if Fodor's arguments succeed, the result is a pyrrhic victory. Where Fodor argues that concepts are innate in the sense of being *unlearnt* (in a technical, and debatable, sense of the word), modern empiricists who avail themselves of advances in machine learning probably won't care. They are likely opposing the more general suggestion that concepts are often, typically, or ever innate in the sense of being genetically hardwired products of our biology that are, thus, "structured in advance of experience" (Marcus 2009: 147). And this – they will maintain – is something they plainly needn't be, as illustrated by the results of neural networks, which seem to abstract new categories from their training data. The upshot is that a nativism that's worth its salt must be motivated quite differently from Fodor's.

With these points in view, I'll now describe a better argument for concept nativism, which avoids these problems. I call this improved argument *The New Riddle of Concept Learning*, partly on account of crude affinities with Goodman's new riddle of induction, but mostly because I like a catchy name.[4]

3.1 The Poverty of the Stimulus

The New Riddle of Concept Learning differs from Fodor's in key respects. In certain respects, The New Riddle motivates a brand of concept nativism that is *stronger* than Fodor's. For instance, it supports the view that concepts are innate in the sense of being *biologically hardwired* and *structured independently of experience,* not merely *unlearnt* in Fodor's technical and idiosyncratic sense of the word. In this way, it motivates a view on which concepts are a product of our biological nature and acquired through normal processes of maturation. To compound matters, it motivates a concept nativism which is stronger than Fodor's in that emphasis is placed on the need to posit both innate concepts *and* associated conceptions of these (i.e., certain specific, structured understandings of these). In this way, it sidesteps the question of whether such innate concepts are atomic – as opposed to (e.g.,) definitions, prototypes, or theories – in a manner that distinguishes The New Riddle from Fodor's nativism concerning primitive concepts. No less importantly, however, there are respects in which The New Riddle commands a nativism that is *weaker* than Fodor's. Most notably, in that it needn't apply to *all* concepts. Thus, I will argue that The New Riddle motivates a view on which numerical concepts, like SEVEN, and folk psychological concepts, like BELIEF, are innate, but not one on which concepts for recent technological innovations, like CARBURETOR, are. I consider these

[4] Emphasis on 'crude': the puzzle that I present emphasizes the descriptive project of explicating how certain concepts *could* be induced, not the normative project of how and when they *should*.

welcome results since they imply that The New Riddle targets concept nativism's critics, in ways which enable us to sidestep controversial questions regarding the structure/format of concepts, all the while ensuring that its conclusions remain sufficiently restricted in scope as to be taken seriously.

To bring the argument into focus, the current section will focus on unpacking one further point of difference: that The New Riddle motivates its nativism by appealing to a fundamentally different explanandum to Fodor's. Where Fodor challenges anti-nativists to explain how novel concepts could be learnt at all, purporting to identify an incoherence in the idea that they might be, The New Riddle challenges anti-nativists to explain how certain concepts could be acquired in the specific ways they are, without accepting that the acquisition process is biologically pre-structured independently of experience. Since this final point is central to appreciating The New Riddle's argumentative force, the present subsection will clarify it, enabling us to situate the argument against Chomsky's well-known conjecture that humans have innate (tacit) knowledge of linguistic grammar.

Chomsky argues that certain knowledge of linguistic grammar is innate because first language learners face a *poverty of the stimulus*: there is not enough information in their environment to explain how they come to understand languages in the specific ways they do. Crucially, the challenge this presents critics is not simply that of explaining how children might learn some possible grammar given their experience, as if this is impossible in the way Fodor takes concept learning to be. Rather, the crux of Chomsky's argument is to explain why children consistently acquire the specific grammars they do, when there are indefinitely many alternatives that prove wholly consistent with their evidence. According to Chomsky, this is particularly hard for non-nativists to explain because – often – the grammars that language learners acquire fail to be simplest or most natural in any obvious respect.

Consider the following examples:

(1) The dog is in the park
(2) Is the dog in the park?
(3) Oscar's dog is barking again now that the storm is raging
(4) Is Oscar's dog barking again now that the storm is raging?

Here we have two examples (1+2 and 3+4) in which a declarative sentence is paired with its corresponding interrogative. As competent speakers of the English language this is easily recognized. But suppose you were tasked with articulating the grammatical rule that governs these transformations, reliably turning English declaratives, like (1) and (3), into their corresponding interrogatives, (2) and (4). What would you say? If you were simply given the isolated examples above, you might be tempted to propose a rule of the following sort: To transform a declarative of the form (1) and (3) into its corresponding interrogative, you simply move the first "is" from the assertion to the front of the sentence. After all, this rule is simple, easily understood, and makes correct predictions in the above examples. It is, thus, the kind of proposal that many of us would be tempted to propose.

Alas, this simple proposal quickly runs into problems. Consider how it applies to (5):

(5) That dog who is barking is a cockerpoo

The result would be:

(6) *Is that dog who barking is a cockerpoo?

As competent speakers of English we immediately recognize that this is ungrammatical. The correct transformation of (5) is, of course:

(7) Is that dog who is barking a cockerpoo?

At this point, your inclination might be to seek some more nuanced rule, which predicts whether it is the first or second 'is' that must be moved to the front of the sentence. There are, in fact, indefinitely many rules of this sort which could be postulated. But, without getting bogged down in the details, it is widely accepted by contemporary linguists that *linear rules* – rules which simply appeal to the ordering of words within a sentence (e.g., whether it is the *first* or *second* or *n*th 'is' that must be moved) – fail to provide psychologically plausible accounts of grammatical competence. Instead, competent language users grasp grammatical rules that are *structure dependent*: rules that are framed in terms of the constituent structure of sentences and the syntactic categories to which their constituents belong. For instance, most linguists accept that the rule we (tacitly) employ when assessing the above transformations is one which dictates moving the *auxiliary verb* from the *main clause* of the sentence to its front.

Notice, however, that structure dependent rules of this sort do not seem to be the simplest or most obvious ways to make sense of linguistic data. When we, amateur linguists, try to explain transformations, like the above, we often posit linear rules. These strike us as simple and natural. But while indefinitely many linear rules could be offered to explain finite subsets of a language, like sentences (1)-(4), the structure dependent rules that ultimately underlie our grammatical competence appear comparatively complicated and opaque, at least pre-theoretically. In the above example, the relevant rule's application requires that we identify main clauses and auxiliary verbs and distinguish these from other syntactic categories (like subordinate clauses). For most educated adults who already speak English, identifying and articulating these categories is difficult – it's the sort of thing we might need Google to clarify and, even then, find unclear without tutelage and examples. The upshot is that an adequate theory of grammatical development needs to explain how language learners arrive at these non-obvious interpretations of their linguistic data, over simpler and more obvious alternatives.

Nativists have an easy explanation for this. They can posit innate knowledge of these otherwise obscure syntactic categories, plus hardwired knowledge that grammatical rules are to be framed in terms of these. More generally, they can posit innate and biologically structured biases and dispositions

to formulate structure dependent grammars when first learning a language. Chomsky urges, however, that it is non-obvious what those who oppose such nativism should say.

One problem is that children receive little or no formal instruction on the need to formulate structure dependent grammars. This is unsurprising since most adults lack an explicit understanding of structure dependency, its centrality to the grammars of natural languages, or of how to characterize the syntactic categories these structure dependent grammars appeal to. To compound matters, there are indefinitely many structure *in*dependent grammars which are consistent with the finite utterances a language learner will have encountered. If contemporary linguists are to be believed, these rules must somehow be rejected or avoided. And, as we can now see, this must occur even though many of these will be consistent with the learner's evidence/experience, and even though these would probably strike them as simpler. Yet, really, all of this is a preamble:

Since structure independent rules strike us as the most natural grammars to posit when interpreting linguistic regularities, we would expect unbiased language learners to *often*, or at least *occasionally,* formulate structure independent, linear rules, when first generating hypotheses about the workings of their language. Yet when linguists pour over corpus data, they find that young children essentially *never* make errors of this sort. For instance, Stromswold (ibid.; 1990; see also Pinker 1994) documented the kinds of grammatical errors that children would make if they failed to distinguish auxiliary and lexical verbs when formulating a grammar. But, when she proceeded to analyze 66,000 auxiliary verb involving utterances, produced by 13 children as young as 11months, she failed to find a single error of this sort.[5] As Stromswold emphasizes, these results would be astronomically unlikely to obtain unless these children were already competent identifying auxiliary verbs and had a pre-existing bias to formulate grammatical rules in terms of these (see also Pinker [1984] and Maratsos [1983]).

Thus, Chomsky argues that to understand linguistic development we must explain why children not only formulate structure dependent grammars when first learning a language, but – also – why they never even seem to consider structure independent grammars. And his suggestion is that it is hard to explain this without positing some innate grasp of/bias towards structure dependency. This is because first language learners face a *poverty of the stimulus*: there is insufficient information in their environment to explain why they would uniformly adopt structure dependent grammars. Their linguistic experience is consistent with countless structure *in*dependent grammars, they receive no obvious instruction on this front, and many of these alternative grammars appear more intuitive. Thus, without some innate knowledge of and/or bias towards structure dependency, we would expect children to often adopt structure independent grammars when learning to speak. Since they never do, Chomsky claims that we have no serious alternative but to posit innate and domain specific knowledge which ensures that children formulate structure dependent grammars in this context.

[5] Stromswold analyzed 14 children's utterances, but one child failed to produce any auxiliary-verbs whatsoever. This illustrates just how early on in linguistic development structure dependent grammars are (unwaveringly) embraced, without any obvious process of trial-and-error.

3.2 From Chomsky to The New Riddle

Chomsky argument is not uncontroversial – for instance, objections have been raised against it in light of recent large language models. We will consider some of these objections in Section 3.3. But, in the first instance, I wish to note that, irrespective of whether Chomsky's argument for *grammatical* nativism is sound, it can be adapted to support a robust brand of *concept* nativism in myriad contested domains, presenting what I am calling The New Riddle of Concept Learning.

In its generic form, The New Riddle of Concept Learning proceeds as follows:

A) When children acquire certain concepts, they construe these in specific ways

But:

B) We would expect children to construe these concepts quite differently if these concepts were not innately pre-structured

This is because:

a. Their experience is consistent with alternative ways of construing these concepts
b. Without innate structuring, these alternatives would strike children as simpler
c. These alternatives are not entertained by children
d. And yet, children are not taught to reject these alternatives

The proposed upshot is that the afflicted concepts need to be innately hardwired and biologically pre-structured to explain why children do not adopt the alternative conceptions referenced in (B).

To frame this argument, we can begin by noting that Rips et al. (2006) and Rey (2014) have criticized influential accounts of how humans learn natural number concepts – like NINE or FIFTY-SEVEN – on related grounds. To this end, both begin from the observation that numerate humans conceive of natural numbers in specific ways. For instance, they appreciate that each natural number is *exactly one larger than its predecessor* and that natural numbers thereby conform to *the successor function*. The upshot is that a complete account of natural number conception must explain how and why humans come to construe the natural numbers in this specific way.

Rips et al. and Rey's concern is that standard accounts of numerical development, on which natural number concepts are learnt, don't explain this. Despite subtle differences, standard accounts maintain that children learn the natural numbers incrementally, first acquiring concepts for small numbers in the subitizing range: ONE, TWO, and THREE. These small number representations are the innate outputs of a small number system, which represents these quantities exactly, without providing the resources to represent larger numbers (Margolis 2021). Or, alternatively, these small number representations are themselves learnt when a parallel individuation system (with a set-size limit of three) enables young children to place small collections (<four) into relations of one-to-one correspondence, such that these can be mapped onto the first three labels in an ordered but initially meaningless linguistic count-list (Carey 2009; Shea 2010; Beck 2017). Either way, standard accounts maintain that, having acquired ONE, TWO, and THREE, children are then poised to discover that each of these numbers is one larger than its predecessor, at which point they make an inductive leap, noting that they could keep generating larger natural numbers that are *one* larger than their predecessor

forever: In the same way that THREE is *one larger* than TWO, they induce that there is a next natural number, FOUR, that is just *one larger* than THREE, and so on into infinity.

Standard accounts of this sort have become standard for a reason. Their emphasis on children's initial mastery of *one, two,* and *three* is supported by at least two empirical observations. Firstly, young infants precisely discriminate collections containing three items or less, while their discrimination of larger quantities is often observed to be characteristically imprecise and ratio-dependent (Feigenson et al. 2004). Secondly, this capacity to precisely discriminate collections containing three items or less is said to be reflected in their initial learning of number words. For instance, in the Give-N task (Wynn 1990; 1992), experimenters measure children's early understanding of words in a count-list they have memorized by asking them to hand over specific numbers of items. The well-known result is that, after a phase in which children show no understanding of the words in their count-list, they become "one-knowers"; reliably passing one item when asked but passing random numbers of items when asked to hand over any larger quantity. Months later, children become "two-knowers", reliably passing one item or two items when asked, and then "three-knowers", reliably passing one, two, or three items when asked, while continuing to pass random numbers of items when asked for larger quantities. Success at these stages is seen to confirm the child's mastery of ONE, TWO, and THREE. But shortly thereafter, something magic happens: Children become "cardinal principle knowers", recognizing that each entry in their count list refers to a value that is exactly one larger than its predecessor. Thus, they begin responding appropriately when asked to pass any number of items referenced in their count list, such that a child who has learnt to recite the numbers up to "ten" will be able to non-accidentally pass exactly *seven* toys or *nine* smarties when asked. On standard accounts, this is because children have made the abovementioned inductive leap, conceptualizing quantities that could not be expressed in terms of their initial unlearnt conceptual repertoire. Since this inductive leap is considered a rational response to evidence – a reasoned generalization from the observed fact that TWO is one larger than ONE and THREE is one larger than TWO – children are seen to have expanded the expressive potential of their conceptual system through a psychological process, apt to be called "learning" even by Fodor's stringent criteria.

A Fodorian might question whether these proposals really explain how children might expand their conceptual repertoire through learning (Fodor 2010). Rips et al. (2006) and Rey (2014) are animated by a quite different worry. Their worry is that while observing that *TWO is one larger than ONE* and *THREE is one larger than TWO* is consistent with the natural numbers conforming to the successor function, it is equally consistent with deviant alternatives. For instance, Rips et al. note that it is equally consistent with a deviant Kripkenstein-esque ordering of the number line, where each natural number is exactly one greater than its predecessor, but only up until nine, at which point the count list returns to zero. Likewise, Rey notes that it is equally consistent with larger numbers conforming to the successor function, such that for any given natural number y which follows immediately after any given natural number x, $y = x + 1$, unless x = 57,453 in which case $y = 2$. But if the child's grasp of ONE, TWO, THREE, and their interrelations, is equally consistent with these (and infinitely many additional) deviant orderings of the natural numbers, Rips et al. and Rey object that standard accounts

have failed to explain why children make the specific inductive leap they do; inducing that each natural number conforms to the successor function, rather than one of the infinitely many deviant alternatives that would prove equally consistent with their prior understanding as three-knowers.

I think that Rips et al. and Rey are raising an important concern. However, the examples they provide obscure the force of the problem. To reject standard accounts of number learning, and motivate a nativist alternative, it is not enough to observe that (A) children come to construe natural numbers as conforming to the successor function, despite (a) their experience/evidence proving equally consistent with deviant alternatives. For if proponents of these standard accounts can identify a non-nativist reason why these deviant alternatives are neglected or rejected by number learners, they can legitimately claim victory; maintaining that they have provided a non-nativist account of how natural number concepts are acquired by normal human beings. This is troubling for Rips et al. and Rey, since it is not hard to come up with plausible reasons of this sort. For instance, a bare appeal to simplicity might suffice: Rips et al. and Rey's examples involve children acquiring the axioms of a deviant arithmetic on which natural numbers conform to the successor function *except in exceptional circumstances*. But, plausibly, the most economical way for a child to encode such deviant orderings, would be by first representing that each natural number conforms to the successor function *and,* then, additionally representing certain exceptions to this general rule (e.g., that if the number in question succeeds nine then it is zero, or that if the number in question succeeds 57,453 then it is two). Since representing a general rule that each natural number conforms to the successor function *plus* exceptions to this general rule is more complicated that simply representing the general rule, it is not unreasonable to suppose that simplicity alone might lead children to reliably favor the hypothesis that natural numbers conform to the successor function *simpliciter*.

The point to note, is that there are deviant orderings of the natural numbers which make Rey and Rips et al.'s point more convincingly, since these deviant orderings *would* be simpler from the perspective of the child who lacks a robust understanding of the successor function.

To illustrate, note that while mainstream accounts of numerical cognition hold that children are, initially, oblivious to exact natural numbers (at least when >3), demanding that these be learnt or otherwise non-innate, they accept that infants possess an innate and early emerging *approximate number system* (ANS). As its name suggests, the ANS is a psychological system which represents – sometimes quite large – numbers, imprecisely or approximately, with its imprecision reflecting conformity to Weber's Law. So, while mature number concepts enable us to precisely enumerate and discriminate two collections based on their number, the ANS's ability to discriminate two numerical quantities is a function of their ratio: the further from 1:1, the better the system performs.

There is inordinate empirical evidence for an innate and early emerging ANS of this sort. In a famous study, Xu and Spelke (2000) presented 6-month-olds with successive collections of dots on a screen. These collections diverged in their cumulative surface area, spatial density, average brightness, and convex hull, but each contained an identical number of dots. The infants would, thus, observe

successive collections of *N* dots, until they became bored, as indexed by significantly decreased looking times when presented with new collections of N dots. At this point, the infants were presented with a final collection, this time containing a novel number of dots. Interestingly, the infants regained interest in these new displays, but only when the number of dots they contained differed from N by a ratio of at least 1:2. Hence, infants who were habituated to successive collections of eight dots would regain interest when presented with collections of 16 or four dots, but not collections of 12. This was so, even though non-numerical confounds had been controlled for.

This result has now been replicated many times (e.g., Lipton & Spelke 2003) with Libertus and Brannon (2010) finding matching levels of performance when 6month-olds were tested on novel change-detection paradigm. Izard et al. (2009) even found analogous abilities in *newborn* infants (under three-days-old) who could, this time, match the approximate number of dots in a seen collection to the number of tones in a heard sequence (see also: de Havia et al. 2014). Aside from speaking to just how early on in development an ANS emerges, cross-modal paradigms of this sort rule out non-numerical confounds as the drivers of the effects; at any rate, no one has ever identified a credible non-numerical confound which could explain how infants match (e.g.,) 12 tones to 12 colored shapes. And because the ANS enables young children and infants to perform basic arithmetic operations, such as appreciating the approximate number of items that would result from two collections being added together, subtracted from one another (McCrink & Wynn 2004), or multiplied (Qu et al. 2021), the ANS appears to be in the business of representing relatively determinate numerical quantities, like *SEVEN* or *SEVENISH* (Clarke 2023). The upshot is that the existence of an innate and congenital ANS is among the best established posits in cognitive science (Clarke & Beck 2021; Dehaene 2011).

Despite widespread agreement that an innate and congenital ANS exists, researchers have neglected to appreciate a dilemma that this poses for standard accounts on which exact natural number concepts are learnt and non-innate. To illustrate, suppose that proponents of these accounts acknowledge the existence of an innate ANS but maintain that, despite its conformity to Weber's Law, the ANS produces exact natural number concepts or representations that conform to the successor function. On this view the ANS might, thus, represent *exact* natural numbers but only apply these to observed collections *imprecisely* such that ANS-based-performance emerges as ratio dependent. The trouble with this (Horn 1 of the dilemma) is that it should be unpalatable to those who hold that exact natural number concepts/representation are learnt, since this seems tantamount to saying that exact natural number representations/concepts are innately endowed products of our ANS.

As an alternative, proponents of these standard accounts can accept the existence of an ANS but deny that an innate ANS produces exact natural number concepts. For reasons we can now appreciate, this is the route that proponents of standard accounts will typically embrace, holding that ANS representations differ from mature and exact number concepts on account of their approximate contents warping number-space in some way or another. Thus, proponents of learning accounts often assert that, unlike learnt number concepts, ANS representations "obscure the successor function" (Carey 2009: 295) by representing the number line as logarithmically compressed (DeWind et al. 2015)

such that 5 is construed as less similar to 6 than 4. But now a further worry arises (Horn 2) for it is hard to see why conformity to the successor function would continue to constitute a simple or natural way to interpret the number line from the perspective of a child first learning to count. Even if a child has noticed that TWO is one larger than ONE and THREE is one larger than TWO, as standard accounts insist, their *only* grasp of larger numerical quantities is now seen to be screaming at them that natural numbers >3 do not conform to the successor function at all.

One might hope to avoid this dilemma by denying the existence of an innate ANS entirely. But this seems desperate since the existence of an innate and congenital ANS is among the best supported posits in all of psychology, anthropology, and neuroscience and its existence is assumed and enthusiastically defended by those who have sought to offer clear and specific accounts of how natural number concepts are learnt (Beck 2017; Carey 2009; Laurence & Margolis 2008; 2024; Shea 2011; Spelke 2017).[6] Likewise, one fails to avoid the dilemma by simply positing additional numerical resources at the learner's disposal. For, unless these numerical resources effectively tell the child how to structure the natural numbers, such that the view collapses into number concept nativism (Horn 1), it remains unclear why conformity to the successor function will appear simple or natural to the child first learning to count (Horn 2). For instance, it won't help to follow Anobile et al. (2016) in supposing that, in addition to an ANS, humans possess a texture density system which facilitates numerical comparisons in conformity with a distinct square root law. For unless this imprecision masks an underlying competence with exact natural numbers, this system will just present yet another deviant structuring of the number line, that children should find more intuitive or obvious that one conforming to the successor function.

A more promising response would involve advocates of standard accounts maintaining that while the ANS distorts the number line, children are somehow taught to avoid the deviant orderings it recommends and, thus, taught to count in ways which conform to the successor function. The idea that children are taught to count is, of course, highly intuitive, and teaching of an appropriate variety might constrain wayward hypotheses about number and mathematics. But even bracketing recent evidence that pre-verbal infants, who do not yet seem to be in a position to be taught to count, already possess the competence to enumerate large numerical quantities precisely, in strict conformity with the successor function (Clarke 2025), the idea that children learn to count by being taught to avoid deviant orderings of the natural numbers is undermined by the very evidence which motivates standard accounts of number learning.

For instance, we've seen that standard accounts of number learning are motivated by the Give-N task, in which one-, two-, and three-knowers reliably give experimenters collections of exactly one, two, or three items, respectively, when asked. What is crucial to recognize, is that when children at these stages

[6] Samuels and Snyder (2024) distinguish two kinds of number content, maintaining that ANS representations represent *cardinalities* – numerical properties of collections – rather that *numbers* – objects to which properties can be attributed. Bracketing concerns with this proposed bifurcation (Beck & Clarke, forthcoming; Clarke & Wellwood MS.), the dilemma under consideration still arises with respect to the cardinality concepts we acquire. Hence, the basic problem remains.

are asked to hand over collections containing larger numbers of items, they pass *random* numbers of items prior to developing into cardinal-principle knowers. Indeed, this is mirrored in other tests of numerical comprehension, like the "What's on this card?" task, in which children are asked to report the number of items on a card (Le Corre et al. 2006). Thus, at no point do they map number words onto approximate numbers of a sort that reflects the ANS's conformity to Weber's Law. Indeed, this much is emphasized by proponents of standard accounts, who argue that children learn to count only after acquiring ONE, TWO, and THREE, using their parallel individuation system (Carey & Barner 2021). So, much as Stromswold observed that children effectively never consider simple and intuitive structure independent grammars when first learning to speak, performance on the Give-N and "What's on this card?" tasks suggests that children effectively never consider that numbers might conform to the deviant structures that should strike them as most intuitive, if their only facility with (non-subitizable) numbers is facilitated by an ANS (and/or texture density system) whose representations distort the successor function.

One might accept this but wonder how the nativist accounts for the fact that (i) children learn the meanings of number words "one", "two", and "three" slowly, incrementally, and in this order, and (ii) why it is that children reliably develop into cardinal-principle knowers only after mastering number words up to "three", specifically. This is pressing since it is these observations which motivate standard accounts of number concept learning which emphasize the importance of a small number or parallel individuation system, with a corresponding set size limit of three (e.g., Carey & Barner 2021).[7]

The nativist can address both points. On (i), children's slow and incremental grasp of the words "one", "two", and "three" is seen to simply reflect difficulties mapping number words onto pre-existing number concepts (Spelke 2017; Margolis 2021); indeed, standard biases in word learning have been seen to predict difficulties of this sort (Clarke 2025). Moreover, an account of this sort would still predict that number words referring to small values are learnt first, because the frequency with which a word referring to a number n is encountered has been shown to robustly follow a $P(n) \propto 1/n^2$ law, such that words for the smallest number are encountered most regularly (Dehaene & Mehler 1992).

On (ii), the nativist can maintain that rather than reflecting an inductively inferred appreciation that natural numbers conform to the successor function, children's immediately transitioning from *three-knower* to *cardinal-principle knower* instead reflects their emerging appreciation that it is a pre-existing grasp of natural numbers that is to be mapped on to number words. Consistent with this latter possibility, but inconsistent with standard accounts of number learning which hold that number words are initially mapped onto a small number or parallel individuation system with a set-size limit of *three* (e.g., Carey & Barner 2021), researchers have found considerable variation in the knower-level that children reach before developing into cardinal-principle knowers. For instance, Krajsci and Fintor

[7] A further motivation for non-nativist accounts of number development pertains to anthropological evidence, where humans who lack exact number words are said to lack exact number concepts (Gordon 2004). This claim deserves careful consideration but doing so is beyond the scope of this paper. Here, I'll simply note that the replicability of such evidence (a) seems to depend on the sympathies of the researchers involved (Butterworth et al. 2008) and (b) seems to dissipate when tests of exact enumeration no longer place excessive demands on memory and mental arithmetic (Izard et al. 2008).

(2023) found that children regularly become four- and five-knowers prior to becoming cardinal-principle knowers. Such variation is precisely what we would expect if children learn to understand number words by incrementally mapping these onto an innately pre-structured and unbounded sequence of natural number concepts.

In sum: Humans acquire natural number concepts in highly specific ways, recognizing that these numbers conform to the successor function and perhaps other axioms of arithmetic. But standard accounts of number concept learning struggle to explain why. This is not simply because there are alternative interpretations of the number line that could be consistent with children's prior understanding and experience (*pace* Rips et al. 2006 & Rey 2014). It is because certain deviant interpretations of the number line would strike children who have not yet *learnt* to conceive of the natural numbers as conforming to the successor function as simpler and more intuitive. Yet rather than being taught to avoid these deviant interpretations, available evidence indicates that children never even consider these. Without some hardwired bias to conceive of the natural numbers as conforming to the successor function – e.g., an innately imbued algorithm for generating arbitrarily large number representations in accord with this function – it is hard to make sense of these results.

3.3 The Challenge from Neural Networks (Redux)

The preceding section presents a challenge for standard accounts of numerical development on which natural number concepts are learnt. One notable feature of these accounts, is that they purport to show how such concepts can be learnt via rational (inductive) processes, thereby engaging Fodor's challenge head on (Beck 2017; Carey 2009; Laurence & Margolis 2002; 2008; Shea 2011). Yet, even if one is persuaded by the problems that The New Riddle poses these accounts, we have seen that modern neural networks may acquire novel representations through a-rational processes, which nevertheless threaten a meaningful concept nativism. Before considering how The New Riddle generalizes to other conceptual domains, it is, thus, pressing to ask: Do recent advances in neural network architectures undermine The New Riddle, as I have presented it?

This is pressing, since it has been claimed that neural networks, like the transformer architectures that large language models (LLMs) like ChatGPT employ, effectively refute Chomsky's claim that there is a poverty of the stimulus in the acquisition of linguistic grammar (Pater et al. 2019; Piantadosi 2023; Warstadt et al. 2019). For, contrary to Chomsky's claim that innate knowledge of grammar is required to make sense of human language learning, critics object that LLMs succeed in producing impressively human-like linguistic outputs when prompted, despite starting out "relatively unconstrained" (Piantadosi 2023: 18), by simply generalizing from linguistic data found on the internet. Indeed, similar points might appear to apply to the acquisition of number concepts. For while I have argued that there is a Poverty of the Stimulus afflicting accounts of number learning, we have seen that convolutional neural networks, which are simply constructed and trained for the purposes of non-numerical object/image classification, can succeed in abstracting numerical contents (Nasr et al. 2019).

Such objections miss the mark. The problem with the abovementioned objections is not – as some would have it – that neural networks must be trained on considerably more data than child learners to reach high levels of performance in the domains of grammar and number. In the linguistic domain, for instance, I am sympathetic Piantadosi's speculation that, in the future, "our methods for training [LLMs] on very small datasets will inevitably improve" (2023: 14; see, for instance, Warstadt et al. 2025; but see also Boeckx 2010: 42-7 for a sense of just how impoverished input can be in human language learners). Similar points might apply to Nasr et al.'s network, which was found to have abstracted numerical contents, but only after being trained on 1.2million labelled images. The real problem is that while these networks achieve interesting levels of performance in the analysis of language and number respectively, the challenge facing anti-nativists is not simply to explain patterns of success; it is to also explain why children do not formulate deviant interpretations of linguistic grammars or the natural numbers, when first learning of these.

It is here that things fall apart for critics of the above sort. There are two problems lurking hereabouts. First, LLMs and networks like Nasr et al.'s make precisely the sorts of errors that an innate grasp of grammar and number seems necessary to prevent. For instance, LLM's grammatical errors often reflect a failure to distinguish auxiliary verbs from other syntactic types (Blevins et al. 2023) and the misapplication of complex linear (structure independent) rules (Yedetore et al. 2023). As we have seen, it is children's failure to *ever* make errors of this sort which supports Chomsky's proposed need for an innate grammar. Importantly, the same applies to the acquisition of number concepts. In Section 3.2, we saw that an account of natural number concept acquisition must account for the fact that children acquire number concepts that conform to the successor function and not – for instance – the ANS's conformity to Weber's Law. But, when Nasr et al.'s network was found to abstract numerical contents, it did so in a manner that was seen to closely resemble ANS imprecision.

Might future neural networks do better in this regard, acquiring exact number concepts that conform to the successor function? Probably. But herein lies a second problem: It is hard to see how a network would do this, reliably and non-accidentally, if a bias or innate tendency to construe numbers in this way was not baked in from the start. Admittedly, we do not yet have a deep understanding of how such biases or knowledge get baked into neural networks (Kodner et al. 2023). But "Ignorance of bias does not imply absence of bias" (Rawski & Heinz 2019). The simple fact is: The reason why we have computer scientists and engineers tinkering with the hyperparameters and architectures of neural networks is that their inbuilt structuring strongly determines *what* the network learns and constrains *how* it does so. If a network were to reliably acquire exact number concepts of a sort that humans employ in the math class, while systematically avoiding the deviant interpretations adopted by Nasr et al.'s network, this would be compelling reason to conclude that its innate structuring dictated this.

3.4 Scope of the Riddle

The New Riddle of Concept Learning supports the view that exact natural number concepts are either innate or innately pre-structured. But, when and where does The New Riddle apply? This is important for two reasons. On the one hand, broader interest in the riddle will reflect the fact that it applies

equally well to a range of contested domains of conceptual development, not just natural numbers. On the other hand, it is important to clarify why it does not entail the implausible conclusion that *all* concepts are innately hardwired products of our biology.

On the first of these points, consider how The New Riddle might be applied to other contested domains. As many readers will be aware, there has been fierce debate as to how children acquire concepts like BELIEF and DESIRE, such that these mental state types can be attributed and used to (e.g.) predict that Sally will search for her marble in the basket where she *falsely believes* it to be (Baron-Cohen et al. 1985). But while disputes as to whether these mental state concepts are innate typically turn on the nitty-gritty empirical details of studies probing young infants' competence in appropriately constructed tasks (Baillargeon et al. 2010), possible confounds in these tasks (Heyes 2014), and their more general replicability (Lavelle 2022), application of The New Riddle confers non-trivial weight to a nativism in this domain, irrespective of how these matters pan out.

As with number concepts, mature folk psychologists deploy concepts like BELIEF and DESIRE in highly specific ways. For instance, we appreciate that if Sally desires her marble, she will likely search for it in the location she falsely believes it to be. But we also appreciate that she will not do this if she also believes that the marble is beside a hungry crocodile; unless, of course, we attribute to Sally a desire to get eaten or a belief that the hungry crocodile is safely caged, etc. Interestingly, large scale meta-analyses suggest that these inferences are remarkably stable across diverse populations (Liu et al. 2008), as does our ability to appreciate complex works of fiction across time. For instance, our facility with belief/desire reasoning presumably needs to be uniform enough that when people peruse The Bible, they can recognize that Judas *wanted* to induce a highly specific *belief* in certain Roman authorities when he kissed Jesus on the cheek, irrespective of whether they be reading this 1900 years ago on the eastern frontiers of the Roman Empire, or in modern day Los Angeles.

Absent innate structuring of BELIEF and DESIRE, it is, however, hard to see how we should account for this. Non-mentalistic, behavioral rules can always be formulated to facilitate the prediction and explanation of observed actions (Povinelli & Vonk 2004). It's, thus, unclear why a creature who lacks mental state concepts would find it simple or natural to ever interpret behavior in mentalistic terms. To compound matters, when creatures do posit hidden mental states as the causes of action, we must contend with the fact that there are indefinitely many models of the mental they could adopt (Butterfill 2017). For both reasons, it is unclear why our specific grasp of mentality would emerge as uniform across time and culture, such that its tenets strike us as natural and intuitive. Indeed, this is particularly troubling if our only innate model of the mental differs markedly from a mature grasp of mentality that must be learnt (Apperly & Butterfill 2009; Butterfill & Apperly 2013; Burge 2018).

In reply, it might be claimed that belief-desire reasoning is *taught* to children and, thus, passed down through generations (Heyes 2018). But it is hard to see how this could be. No one has ever managed to articulate the rules or principles that govern mature belief-desire reasoning (Maibom 2003). In fact, doing so seems to be impossible because no finite number of belief and/or desire attributions

rationally dictates any given behavior. Hence, Lewis despaired of folk psychology that while "We can tell which particular predictions and explanations conform to its principles… we cannot expound those principles systematically" (1994). But if these principles cannot be expounded, it is hard to see how they might be taught in sufficient detail that they could be reliably grasped in uniform ways. This is to say nothing of the fact that while neural networks have been found to succeed on certain litmus tests of belief reasoning in children, such as verbal false belief tasks, their "performance [on these tasks] is not robust against trivial alterations to stimuli" (Pi et al. 2024). Thus, The New Riddle suggests that mental state concepts need to be innate, in much the same way as natural number concepts.

We could keep going. *Prima facie*, The New Riddle generalizes to the acquisition of moral concepts that feature in judgements that are remarkably stable across diverse populations (Mikhail 2002) and comparative concepts, like MORE and MOST (Clarke & Wellwood, forthcoming), which are deployed in opaque yet distinctive ways, even under conditions where their truth conditions do not differ (Knowlton et al. 2021).

At the same time, The New Riddle does not imply the innateness of *all* concepts. This is important, since the reader might now be wondering *why not?* Suppose, for example, that you are in the position of a sheltered philosopher, learning or otherwise acquiring the concept CARBURETOR. This might begin when, upon reading Jerry Fodor, you are told that this concept is innate. At this point, you will have little sense of exactly what a carburetor is. But given general concerns about the inscrutability of reference, one might harbor a deeper worry: that further experience could never settle this. For instance, our sheltered philosopher might proceed to ask Google what a carburetor is and be told that it is a device that mixes air and fuel. But, of course, the worry remains that this is unsatisfactory as a definition – for a start, it fails to distinguish carburetors from fuel injection systems – and absent some such definition, which appropriately specifies the necessary and sufficient conditions on something's being a carburetor (something which seems unlikely to be forthcoming – Fodor 1975), coupled with some pre-existing capacity to appreciate this definition, it may be hard to see how there could avoid being a poverty of the stimulus, such that one's experience will always be consistent with a myriad of deviant interpretations of the concept (Rey 2014). Are we not, then, stuck concluding that CARBURETOR is innate in much the same way as SEVEN or BELIEF appear to be?

I think not. In the case of SEVEN and BELIEF, I argued that the acquisition process is not or cannot be taught. But our conception of CARBURETOR must be, with explicit instruction playing a crucial role in narrowing the space of possible interpretations. This might involve Googling of the abovementioned sort. Alternatively/additionally, it might involve the testimony of an expert, charged with teaching car maintenance. Of course, for this guidance to help, there will presumably still need to be some kind of innate structuring, which constrains how it gets interpreted (Gleitman 1990; Csibra & Gergely 2009). But there is no reason to assume that this innate structuring is specific to the domain of carburetors or car parts (as opposed to – say – artefacts more generally); and when considered alongside the aforementioned instruction it is non-obvious why there would remain a poverty of the stimulus to be overcome. This is because, I see no obvious reason to suppose that acquiring a

sophisticated conception of CARBURETOR involves subjects systematically and uniformly neglecting to consider simpler construals, once the abovementioned considerations are brought into focus. So, while The New Riddle confers substantial weight to the conjecture that contested concepts, like SEVEN and BELIEF are innately hardwired and, hence, biologically pre-structured independently of experience, the argument does not require that things spiral out of control, implying that *all* concepts are innate and biologically pre-structured in this way.

4. Conclusion

Philosophical discussions of concept nativism have often been dominated by Fodor's arguments for "mad-dog nativism" and critics' attempts to rebut these. The present treatment suggests that this is a distraction. Advances in artificial intelligence show that novel concepts might be flexibly acquired from experience. And while Fodor has the resources to accommodate this, the result is a pyrrhic victory: it leaves Fodor defending a brand of concept nativism that is consistent with the ambitions of hard-lined empiricists. For this reason, I suggest that concept nativists are better served by *The New Riddle of Concept Learning*. At its core, The New Riddle proceeds, not by trying to isolate an apparent incoherence in the idea that novel concepts might be learnt from experience, but rather, by asking how we can account for the specific ways in which certain concepts are construed when first acquired. Given that The New Riddle withstands challenges that arise from recent advances in modern artificial intelligence and theories of conceptual development, I submit that The New Riddle of Concept Learning should take a central role in future debates regarding the origins of human concepts.

References:

1. Anobile, G., Castaldi, E., Turi, M., Tinelli, F., & Burr, D. C. (2016). Numerosity but not texture-density discrimination correlates with math ability in children. *Developmental Psychology, 52*(8), 1206-1216.
2. Antony, L. (2020). Not rational, but not brutely causal either: A response to Fodor on concept acquisition. *Theoria,* 35(1): 45-57.
3. Apperly, I. A., & Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review, 116*(4), 953–970.
4. Baillargeon, R., Scott, R. M., & He, Z. (2010). False-belief understanding in infants. *Trends in cognitive sciences*, *14*(3), 110–118.
5. Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"?. *Cognition*, *21*(1), 37–46.
6. Beck J. (2017). Can bootstrapping explain concept learning?. *Cognition*, *158*, 110–121.
7. Berkeley, G. (1710/1975). *A Treatise Concerning the Principles of Human Knowledge*. In M. R. Ayers (ed.), *G. Berkeley, Philosophical Works*, pp. 61–127. Totowa, NJ: Rowman & Littlefield.
8. Blevins, T., Gonen, H., & Zettlemoyer, L. (2023). Prompting language models for linguistic structure. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 6649-6664). Association for Computational Linguistics.
9. Boeckx, C. (2010). *Language in Cognition: Uncovering mental structures and the rules behind them.* Wiley Blackwell.
10. Buckner, C. (2018) Empiricism without magic: transformational abstraction in deep convolutional neural networks. *Synthese* 195, 5339–5372.
11. Buckner, C. (2023) *From deep learning to rational machines.* Oxford University Press.

12. Butterfill, Stephen A. (2017). Tracking and Representing Others' Mental States. In *Routledge Companion to the Philosophy of Animal Minds* (pp. 269--279). Routledge.
13. Butterfill, S.A. and Apperly, I.A. (2013), How to Construct a Minimal Theory of Mind. *Mind & Language*, 28: 606-637.
14. Burge, T. (2018). Do infants and nonhuman animals attribute mental states? *Psychological Review, 125*(3), 409–434.
15. Carey, S. (2009). *The Origin of Concepts*. Oxford University Press.
16. Churchland, P. (1986). Neurophilosophy. Cambridge, MA: MIT Press.
17. Clark, A. (1993). *Associative engines: Connectionism, concepts, and representational change*. The MIT Press.
18. Clarke, S. (2023). Compositionality and constituent structure in the analogue mind. *Philosophical Perspectives*, 37: 90–118.
19. Clarke, S. (2025). Number nativism. *Philosophy and Phenomenological Research*, 110, 226–252.
20. Clarke, S., & Beck, J. (2021). The number sense represents (rational) numbers. *Behavioral and brain sciences*, *44*, e178.
21. Clarke, S. & Wellwood, A. (forthcoming). Three brands of concept nativism. *Analysis*.
22. Clatterbuck, H. & Gentry, H. (2025). Learning incommensurate concepts. Synthese 205 (3):1-36.
23. Csibra, G., & Gergely, G. (2009). Natural pedagogy. *Trends in cognitive sciences*, *13*(4), 148–153.
24. Dehaene, S., & Mehler, J. (1992). Cross-linguistic regularities in the frequency of number words. *Cognition*, *43*(1), 1–29.
25. de Hevia, M. D., Izard, V., Coubart, A., Spelke, E. S., & Streri, A. (2014). Representations of space, time, and number in neonates. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(13), 4809–4813.
26. DeWind, N. K., Adams, G. K., Platt, M. L., & Brannon, E. M. (2015). Modeling the approximate number system to quantify the contribution of visual stimulus features. *Cognition*, *142*, 247–265.
27. Elbourne, P. (2011). *Meaning: A Slim Guide to Semantics*. Oxford: Oxford University Press.
28. Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in cognitive sciences*, *8*(7), 307–314.
29. Fodor, J. A. (1974). Special Sciences (Or: The Disunity of Science as a Working Hypothesis). *Synthese*, *28*(2), 97–115.
30. Fodor, J. (1975) *The Language of Thought*, New York: Thomas Y. Crowell.
31. Fodor, J. (1981) *Representations*, Cambridge, MA: MIT Press.
32. Fodor, J. (1994) *The Elm and the Expert*, Cambridge, MA: MIT Press.
33. Fodor, J. (1997). Special Sciences: Still Autonomous After All These Years. *Philosophical Perspectives*, *11*, 149–163.
34. Fodor, J. (2008). *LOT 2: The Language of Thought Revisited*. Oxford: Oxford University Press.
35. Fodor, J. (2010). Woof, Woof. Review of *The Origin of Concepts* by Susan Carey, *The Times Literary Supplement*, October 8: pp. 7–8.
36. Fodor, J., & McLaughlin, B. P. (1990). Connectionism and the problem of systematicity: Why Smolensky's solution doesn't work. *Cognition, 35*(2), 183–204.
37. Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition, 28*(1-2), 3–71.
38. Gelman, R., & Gallistel, C. R. (1978). *The childs understanding of number*. Cambridge, MA: Harvard University Press.
39. Gleitman, L. (1990). The structural sources of verb meanings. *Language acquisition*, *1*(1), 3-55.
40. Heyes C. (2014). False belief in infancy: a fresh look. *Developmental science*, *17*(5), 647–659.
41. Heyes, C. (2018). *Cognitive gadgets: The cultural evolution of thinking*. Harvard University Press.
42. Hume, D. (1739/1978). *A Treatise of Human Nature*, ed. L. A. Selby-Bigge, revised P. H. Nidditch. Oxford: Oxford University Press.
43. Izard, V., Sann, C., Spelke, E. S., & Streri, A. (2009). Newborn infants perceive abstract numbers. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(25), 10382–10385.
44. James, W. (1890). *The Principles of Psychology*, Vol. 1. Mineola, NY: Dover Publications Inc.

45. Kelleher, J.D. (2019). *Deep learning*. MIT Press.
46. Knowlton, T., Hunter, T., Odic, D., Wellwood, A., Halberda, J., Pietroski, P., & Lidz, J. (2021). Linguistic meanings as cognitive instructions. *Annals of the New York Academy of Sciences*, *1500*, 134–144.
47. Kodner, J., Payne, S., & Heinz, J. (2023). *Why linguistics will thrive in the 21st century: A reply to Piantadosi,* arXiv: 2308.03228).
48. Krajcsi, A., & Reynvoet, B. (2024). Miscategorized subset-knowers: Five- and six-knowers can compare only the numbers they know. *Developmental science*, *27*(1), e13430.
49. Laurence, S., & Margolis, E. (2002). Radical concept nativism. *Cognition*, *86*(1), 25–55.
50. Margolis, E., & Laurence, S. (2008). How to learn the natural numbers: inductive inference and the acquisition of number concepts. *Cognition*, *106*(2), 924–939.
51. Margolis, E., & Laurence, S. (2024). *The Building Blocks of Thought: A Rationalist Account of the Origin of Concepts*. Oxford: Oxford University Press.
52. Lavelle, J. S. (2022). When a crisis becomes an opportunity: The role of replications in making better theories. *The British Journal for the Philosophy of Science, 73*(4), 965-986.
53. Le Corre, M., Van de Walle, G., Brannon, E. M., & Carey, S. (2006). Re-visiting the competence/performance debate in the acquisition of the counting principles. *Cognitive psychology*, *52*(2), 130–169.
54. Lewis, D. (1994). Reduction of Mind. In S. Guttenplan (ed.), *A Companion to Philosophy of Mind*, Oxford: Blackwell, pp. 412–31.
55. Lipton, J. S., & Spelke, E. S. (2003). Origins of number sense. Large-number discrimination in human infants. *Psychological science*, *14*(5), 396–401.
56. Liu, D., Wellman, H. M., Tardif, T., & Sabbagh, M. A. (2008). Theory of mind development in Chinese children: a meta-analysis of false-belief understanding across cultures and languages. *Developmental psychology*, *44*(2), 523–531.
57. Locke, J. (1690/1975). *An Essay Concerning Human Understanding*, ed. P. H. Nidditch. Oxford: Oxford University Press.
58. Maibom, H. (2003), The Mindreader and the Scientist. *Mind & Language*, 18: 296-315.
59. Maratsos, M. (1983). Some current issues in the study of the acquisition of grammar. In P. H. Mussen (Ed.), *Handbook of child psychology: Formerly Carmichael's manual of child psychology*. Wiley.
60. Marcus, G. (2009), How *Does* the Mind Work? Insights from Biology. Topics in Cognitive Science, 1: 145-172.
61. Margolis, E. (2020). The Small Number System. *Philosophy of Science*, 87(1), 113–134.
62. McCrink, K., & Wynn, K. (2004). Large-number addition and subtraction by 9-month-old infants. *Psychological science*, *15*(11), 776–781.
63. Mikhail, J. (2007). Universal moral grammar: Theory, evidence and the future. *Trends in Cognitive Sciences, 11*(4), 143–152.
64. Mill, J. S. (1882). *A System of Logic, Ratiocinative and Inductive*. London: Longmans, Green, and Co.
65. Nasr, K., Viswanathan, P., & Nieder, A. (2019). Number detectors spontaneously emerge in a deep neural network designed for visual object recognition. *Science Advances, 5*(5), eaav7903.
66. Pater, J. (2019). Generative linguistics and neural networks at 60: Foundation, friction, and fusion. *Language*, 95(1):e41–e74.
67. Perfors, A., Tenenbaum, J. B., Griffiths, T. L., & Xu, F. (2011). A tutorial introduction to Bayesian models of cognitive development. *Cognition*, *120*(3), 302–321.
68. Pi, Z., Vadaparty, A., Bergen, B. K., & Jones, C. R. (2024). Dissecting the Ullman variations with a SCALPEL: Why do LLMs fail at trivial alterations to the false belief task? arXiv: 2308.03228

69. Piantadosi, S. T. (2023). Modern language models refute Chomsky's approach to language. *From fieldwork to linguistic theory: A tribute to Dan Everett*, *15*, 353-414.
70. Pinker, S. (1984). *Language learnability and language development.* Harvard University Press.
71. Pinker, S. (1994). The language instinct. New York: William Morrow.
72. Povinelli, D.J. and Vonk, J. (2004), We Don't Need a Microscope to Explore the Chimpanzee's Mind. *Mind & Language*, 19: 1-28.
73. Putnam, H. (1988). *Representation and reality*. Cambridge, MA: MIT Press.
74. Qu, C., Szkudlarek, E., & Brannon, E. M. (2021). Approximate multiplication in young children prior to multiplication instruction. *Journal of Experimental Child Psychology, 207,* 105116.
75. Rawski, J. and Heinz, J. (2019). No Free Lunch in Linguistics or Machine Learning: Response to Pater. Language, 95(1):e125–e135.
76. Reid, T. (1785/2002). *Essays on the Intellectual Powers of Man*, ed. D. Brookes & K. Haakonssen. Edinburgh: Edinburgh University Press.
77. Rey, G. (2014). Innate *and* learned: Carey, Mad Dog nativism, and the poverty of stimuli and analogies (yet again). *Mind & Language, 29*(2), 109–132.
78. Rousselle, L., & Vossius, L. (2021). Acquiring the Cardinal Knowledge of Number Words: A Conceptual Replication. *Journal of Numerical Cognition*, *7*(3), Article e7029.
79. Rips, L. J., Asmuth, J., & Bloomfield, A. (2006). Giving the boot to the bootstrap: how not to learn the natural numbers. *Cognition*, *101*(3), B51–B60.
80. Russell, B. (1912). *The Problems of Philosophy*. Oxford: Oxford University Press.
81. Samuels, Richard (2004). Innateness in cognitive science. *Trends in Cognitive Sciences* 8 (3):136-141.
82. Shea, Nicholas (2011). New concepts can be learned: Susan Carey, The Origin of Concepts, Oxford University Press, Oxford, 2009. *Biology and Philosophy* 26 (1):129 - 139.
83. Spelke, E. S. (2017). Core knowledge, language, and number. *Language Learning and Development, 13*(2), 147–170.
84. Sterelny, K. (1989). Fodor's Nativism. *Philosophical Studies*, *55*(2), 119–141.
85. Stromswold, K. J. (1990). *Learnability and the acquisition of auxiliaries* [Doctoral dissertation, Massachusetts Institute of Technology]. DSpace@MIT. http://hdl.handle.net/1721.1/13715
86. Warstadt, A., Singh, A., and Bowman, S. R. (2019). Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.
87. Warstadt, A., Mueller, A., et al. (2023). Findings of the BabyLM Challenge: Sample-Efficient Pretraining on Developmentally Plausible Corpora. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 1–34, Singapore. Association for Computational Linguistics.
88. Weiskopf, D. (2008). The origins of concepts. *Philosophical Studies,* 140:359-84.
89. Wynn, K. (1990). Children's understanding of counting. *Cognition, 36*(2), 155–193.
90. Xu, F., & Spelke, E. S. (2000). Large number discrimination in 6-month-old infants. *Cognition*, *74*(1), B1–B11.
91. Yedetore, A., Linzen, T., Frank, R., & McCoy, R. T. (2023). How poor is the stimulus? Evaluating hierarchical generalization in neural networks trained on child-directed speech. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 9370–9393). Association for Computational Linguistics.