The New Riddle of Concept Learning*

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

<Draft as of May 26, 2026 – comments and hate mail welcome>

Concept nativists often argue that (certain) concepts are innate because their acquisition would – otherwise – be impossible for the organisms in question, given their situation and/or cognitive resources. This leaves nativists vulnerable to surprising developments in AI and cognitive science which suggest that impossible-seeming processes of non-nativist concept acquisition can be achieved. The present paper develops an alternative argument for concept nativism that is insulated from these developments. Drawing on Chomsky, *The New Riddle of Concept Learning* challenges anti-nativists to not only explain how target concepts get acquired but, also, why it is that deviant interpretations of these concepts neglect to be adopted, even when they would be simpler and more natural from the learner's perspective (absent innate pre-structuring). This challenge is found to motivate concept nativism in domains like natural number and theory of mind, to evade objections from AI and developmental science, and to avoid overgeneralizing by implying that all concepts are innate. The result is a powerful, yet underappreciated, motivation for a bold, yet plausible, brand of concept nativism.

Keywords: Rationalism, Innateness, Concepts, Cognitive Development, Artificial Intelligence

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 holds that many concepts must be innately bestowed upon us by God or biology, recent advances in artificial intelligence and cognitive development have been seen to rebut these suggestions. For instance, neural networks are said to abstract novel contents (Buckner 2018) in ways that vindicate the contrasting empiricist conviction that experience-driven learning expands our conceptual horizons without innate pre-structuring of a relevant sort (Rogers & McClelland 2004; Dantzig et al. 2011; Buckner 2023). Meanwhile, theories of cognitive development have sought to show that new concepts can be learnt from experience, thereby expanding the expressive potential of our conceptual system (Carey 2009; Perfors et al. 2011).

The present paper aims to temper this rising wave of anti-nativist enthusiasm. To this end, I develop an argument that I call *The New Riddle of Concept Learning*. This argument builds upon insights from

* For helpful (often detailed) comments and discussion, I would like to thank Luca Barlassina, David Barner, Cameron Buckner, Zoe Drayson, Gabe Dupre, Veronica Gómez Sánchez, Steven Gross, Kevin Lande, Deniz Rudin, Henry Schiller, Alexis Wellwood, and an audience at UC Davis.

Chomskian linguistics, where the need to explain negative evidence looms large. To wit: Chomsky emphasizes that empiricist accounts of language acquisition must not only explain the (positive) fact that natural language grammars are acquired by first language learners; they must also explain the (negative) fact that alternative grammars neglect to be acquired, even when these would be simpler from the learner's perspective. After explaining why this has motivated a grammatical nativism, I show that an analogous line of reasoning provides strong motivation for concept nativism in contested domains, like that of natural number and theory of mind. Along the way, I show that findings from machine learning fail to undermine these conclusions, and I proceed to show that the argument does not overgeneralize, implying that *all* human concepts (e.g., CARBURETOR) are innate (c.f., Fodor 1980). The result is a powerful, yet underappreciated, motivation for a robust, yet plausible, brand of concept nativism. But, first, let me explain why The New Riddle is a welcome addition to the nativist's arsenal, contrasting it with a crop of better-known arguments in the vicinity.

1. The Landscape of Nativist Arguments

Arguments for concept nativism take many forms (Laurence & Margolis 2024, chapters 8-16). Nevertheless, they regularly proceed by emphasizing the (positive) fact that organisms acquire some target concept, *C*, before seeking to establish that they could not have possibly done so through unstructured learning. There is often much to be said for these arguments. However, they face familiar challenges, and I worry that overemphasizing arguments of this sort has left nativists in an unduly precarious position given recent advances in AI and developmental science.

Consider a couple of representative examples: One line of nativist argumentation alludes to experimental results suggesting that young infants possess a conceptual competence in domains such as objecthood, number, agency, and space (Spelke 2022). From this, nativists proceed to argue that such competencies appear *too* early to have been acquired through experience-driven learning (e.g., Baillargeon 1987). In other words, they conjecture that it would be impossible (or otherwise implausible) for young infants to have landed upon these conceptual competencies without innate pre-structuring in the target domains. To my mind, this is an attractive suggestion. Nevertheless, the concern has always been that it is notoriously difficult to say quite how much exposure to a given domain would be needed before a concept's learnt acquisition becomes plausible (e.g., Munakata et al. 1997). Indeed, this concern has fangs given recent work in computer science, which shows that neural networks can achieve high levels of performance when trained on smaller datasets than previously expected (Gunasekar et al. 2023; Warstadt et al. 2023; Sachdeva et al. 2024) and that they may abstract novel contents that are neither explicitly coded for in their training data nor directly targeted by their training algorithms (Nasr et al. 2019; see Section 4).

A second line of nativist argumentation, associated with Jerry Fodor (1975; 1980; 2008), is quite different, but faces a structurally analogous concern. Unlike the above argument, which turns on empirical results, Fodor argues that concept learning is a confused idea. The proposed upshot is that *all* primitive lexical concepts (including CARBURETOR and QUARK – Fodor 1975; 1980), and even complex concepts (like GREEN OR BROWN FAIRY CAKES – Fodor 2008) must be innate in the sense of being unlearnt. This is a bold claim! But without launching into a full discussion of Fodor's reasoning, it is akin to the above argument in that its dialectical force turns on a premise that acquiring target concepts (in this case, *all* primitive [Fodor 1980] and complex [2008] concepts) would be impossible via non-nativist means. For this reason, it is – again – open to rebuttal by research which shows that the purportedly impossible processes of acquisition can be achieved. Thus, it is a concern that Fodor's critics have devoted considerable energy to arguing that Fodor's reasoning turns on mistaken claims regarding the structure of concepts (e.g., that lexical concepts are atomic and

semantically unstructured – c.f., Prinz 2002, but see Fodor 2008) and the nature of learning (that learning requires the rational projection and confirmation of hypotheses – c.f., Laurence & Margolis 2024). Worse still, theories in developmental psychology (Carey 2009; Carey & Barner 2019), and the philosophy thereof (Laurence & Margolis 2002; Carey 2009; Shea 2011; Beck 2017) have sought to demonstrate that concept learning, as Fodor understands this, *is* possible, *despite* Fodor's argumentation. All of this leaves Fodor's radical conclusion very much in doubt.

There is, of course, more to be said about these arguments, the specific challenges they face, and other arguments in the vicinity. Nevertheless, the overarching moral is that nativist arguments which turn on claims that an organism could not possibly/plausibly acquire concepts via non-nativist means leave nativists in a precarious position. For arguments of this sort always run the risk of rebuttal by current or future developments in machine learning, developmental psychology, or related fields, which show how these allegedly impossible processes of concept acquisition can be achieved. This is particularly troubling at this unprecedented time of scientific and technological advancement, where developments in AI and cognitive science seem to regularly unearth surprising possibilities.

With this in view, The New Riddle of Concept Learning should be seen as belonging to a rather different crop of nativist argumentation. Rather than emphasizing the (positive) fact that organisms acquire certain concepts, before proceeding to argue that this acquisition would be impossible absent innate pre-structuring, The New Riddle challenges anti-nativists to explain why it is that humans converge on the specific concepts they do, under the specific interpretations that they do, given that alternative interpretations of their concepts (i) are equally consistent with their evidence, and (ii) would often be simpler for them to adopt absent innate pre-structuring. In other words, it challenges anti-nativists to explain both the positive fact that certain concepts are acquired, *and* the negative fact that other concepts (that we would expect them to adopt, absent innate pre-structuring) are not. My suggestion will be that this latter point, regarding negative evidence, motivates concept nativism in myriad contested domains, like natural number and belief-desire reasoning, and that it does so in a way that is unlikely to be undermined by future developments in AI or cognitive science.

Before articulating this argument in detail, three clarifications are in order:

- First, to pose a forceful challenge to anti-nativists, The New Riddle does not merely require that alternative interpretations of the concepts in question are *conceivable* given the subject's experience. It also requires that these alternative interpretations be *simpler* or *more natural* for them to adopt given their pre-existing cognitive resources. As we will see, this is an empirical claim that has been glossed over when related arguments have been made in support of concept nativism (Rips et al. 2006; Rey 2014). However, we will soon see that it serves an important role in the argument. Thus, I devote considerable space to arguing for this premise in each conceptual domain I consider.
- Second, the argument that I develop here targets the innateness of concepts and what we might call our *canonical understanding* of these – the inferential and structural commitments that constitute our mature grasp of them. While much of my discussion will be devoted to motivating the innateness of these canonical understandings in the target domains I consider, I take this to imply the innateness of the concepts themselves, even if the concepts in question be atomic and unstructured, such that these canonical understandings be additional or orthogonal to the concepts themselves (Fodor 1998). This is because, even on such a view, it is hard to see how one could possess an understanding (canonical or otherwise) that *C behaves like thus and so*, without possession of *C* itself (ibid.). At the same time, and unlike some arguments for concept nativism (Fodor 1975; 1980), the argument that I advance does not

presuppose conceptual atomism. It allows that our canonical understandings be constitutive of the concepts in question (e.g., Peacocke 1992) and/or that they serve as an integral part of the machinery by which concepts facilitate categorization (e.g., Smith et al. 1984).

- Third, the argument that I make is *conditional*: it inherits its basic structure from Chomsky's Poverty of the Stimulus argument, and some of its force depends on that framework being broadly correct. I will address this dependency directly in Section 4, when I argue that challenges to this style of argument – deriving from large language models and convolutional neural networks – backfire, since such systems make precisely the kinds of errors that an innate pre-structuring is needed to prevent. To appreciate this, I will begin by considering Chomsky's argument at length (Section 2) such that the uninitiated will be able to see how its basic structure can be marshalled in support of concept nativism in domains like number (Section 3) and theory of mind (Section 5). [Those who are already familiar with Chomsky's arguments can skip ahead to Section 3, where the argument for concept nativism really gets going.]

2. Chomsky's Poverty of the Stimulus

Before articulating The New Riddle, it is instructive to note that Chomsky (1965; 1980) has mounted a related argument, according to which certain grammatical knowledge is innate on the grounds that first language learners face a *poverty of the stimulus*: i.e., because there is simply not enough information in their environment to explain how they would otherwise come to understand languages in the specific ways they do. Crucially, the challenge this presents critics is not simply that of explaining how children learn some possible grammar given their experience, as if this were impossible in the way nativists sometimes take concept learning to be. The crux of the argument is, rather, to explain why children consistently acquire the specific grammars they do, when there are indefinitely many alternative grammars that would be wholly consistent with their evidence. Indeed, this is particularly hard for non-nativists to explain because – often enough – the grammars that language learners acquire fail to be simplest or most natural in any obvious respect.

Take 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 English, 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?

With this pointed out, 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 our grammatical competence. Instead, competent language users are said to grasp grammatical rules that are *structure dependent*: rules framed in terms of the constituent structure of sentences and the syntactic categories to which constituents belong. For instance, most linguists hold that the rule we employ when assessing the above examples is one which dictates moving the *auxiliary verb* from the *main clause* to the front of the sentence (see Laurence & Margolis 2001 for discussion).

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 time of 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 them. More generally, they can posit innate and domain specific biases to formulate structure dependent grammars when first learning a language. By contrast, Chomsky notes that it is non-obvious what those who oppose such nativism can 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 any 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 encounter. 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 presumably 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. But when linguists pour over corpus data, they find that young children essentially *never* make errors of this sort. For instance, Laurence and Margolis (2001) note that Stromswold (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. Yet, 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![1] These results would be astronomically unlikely to obtain unless 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]).

So, in sum: Chomsky argues that understanding linguistic development requires us to 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 this is hard to explain without positing some innate and pre-structured grasp of structure dependency. This is because first language learners face a *poverty of the stimulus*: there is insufficient information in their environment to explain why children would uniformly adopt structure dependent grammars. Their linguistic experience is consistent with structure independent grammars, they receive no obvious instruction on this front, and many of these alternative grammars seem simpler and more intuitive. Without some innate grasp of structure dependency, we would expect children to occasionally or often adopt structure independent grammars when learning to speak. Since they effectively *never do*, Chomsky's submits that we have no serious alternative but to posit innate and domain specific tacit knowledge which ensures that children formulate structure dependent grammars in this context.

3. From Chomsky to Concepts

Chomsky's argument is controversial (see Berwick et al. 2011 for sustained defense). In Section 4, we will return to consider certain objections that have been raised against it given recent AI. But, before we do so, I wish to note that – irrespective of whether Chomsky's argument for *grammatical* nativism is sound – it can be adapted to motivate a *concept* nativism in myriad contested domains. The resulting argument is what I call *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.[2]

In its generic form, The New Riddle of Concept Learning riffs on Chomsky's argument (above), and proceeds as follows:

A) When children acquire certain concepts, they construe these in specific ways, consistently adopting certain canonical understandings of these

But:

[1] 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 apparent process of trial-and-error.

[2] 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 specifying how/when they should.

B) We would expect children to construe these concepts quite differently, and to adopt quite different canonical understandings, 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. And yet, children are not taught to reject these alternatives

The upshot is that the afflicted concepts, and our canonical understanding of these, need to be innately hardwired or pre-structured to explain why children do not adopt the alternative conceptions referenced in (B). This is for much the same reason that Chomsky thinks an innate grasp of grammar is required to explain why children do not entertain linear rules when first learning language.

To see how this argument works, and how it applies to a contested case study, note that Rips et al. (2006) and Rey (2014) criticize influential accounts of number learning – like NINE or FIFTY-SEVEN – on related grounds. 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 numerical development must explain how humans come to construe natural numbers in this highly specific way.

Rips et al. and Rey's concern is that standard accounts of numerical development, familiar from developmental science, fail to do 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 (Barner 2017; 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; Spelke 2017). Either way, standard accounts maintain that, having acquired ONE, TWO, and THREE, children are 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 (Carey & Barner 2019).[3]

Standard accounts of this sort have become standard for a reason. 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 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

[3] This may be a simplification. On some versions of the standard view, the transition from ONE, TWO, and THREE to *successor function* proceeds via an intermediate stage at which children possess a counting algorithm which mirrors the successor function but is limited to entries in their count-list and – thus – not fully generative (e.g., Schneider et al. 2021). Nothing I'll say turns on this, so I'll put it to one side.

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 while continuing to pass 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 still passing 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" can 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 conceptual repertoire. Since this inductive leap is considered a rational response to evidence – a reasoned generalization from the 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 on stringent accounts of what this requires (e.g., Fodor 1980; 2008).

Of course, some nativists question whether these proposals really explain how children might expand their conceptual repertoire through learning, or if they even hint at how this might be possible (Fodor 2010). Rips et al. (2006) and Rey (2014) are animated by a different worry. Their concern 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 an infinite swathe of deviant alternatives. For instance, Rips et al. note that it is equally consistent with a Kripkensteinian 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 induction that they do; inducing that each natural number conforms to the successor function, rather than one of the infinite deviant alternatives that would be equally consistent with their prior understanding as three-knowers.

I believe that Rips et al. and Rey have raised an important concern. However, I also believe that the specific examples they provide obscure the crux of the issue. To reject standard accounts of number learning, and motivate a nativist alternative, it is not enough to observe that children come to construe natural numbers as conforming to the successor function, despite 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 get neglected by number learners, they can legitimately claim victory; maintaining that they have provided a non-nativist explanation for how natural number concepts are acquired by normal humans. 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 all involve children acquiring the axioms of a deviant arithmetic on which natural numbers conform to the successor function *except in exceptional circumstances*. But, the most economical way for a child to encode such deviant orderings, would perhaps be by first representing that each natural number conforms to the successor function *and*

additionally representing exceptions to this 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 (see, e.g., *the elsewhere condition* in Anderson 1969, Halle 1997, Brown & Hippisley 2012, and Yang 2016 for an empirically motivated cognitive architecture which presupposes this). Since representing a general rule that each natural number conforms to the successor function *plus* exceptions to this seems more complicated than 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*. So, much as Chomsky's argument for grammatical nativism would not be so vexing if it merely challenged non-nativists to explain why language learners arrive at the simplest and most intuitive interpretations of linguistic grammar (despite the conceivability of deviant alternatives), Rey and Rips et al. have failed to present a clear or decisive challenge to their opponents.

There are, however, deviant orderings of the natural numbers which make Rey and Rips et al.'s point more convincingly. This is because these deviant orderings *would* be simpler from the perspective of the child who lacks a robust understanding of the successor function.

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 evidence for an 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 were effectively 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 a 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 these results; 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 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*, and not simply *MORE* or *LESS* (Clarke 2023). An 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 it poses for standard accounts on which exact natural number concepts are learnt. 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 represents *exact* natural numbers, but it applies these to observed collections *imprecisely* such that ANS-based-performance emerges as ratio dependent. The trouble with this (Horn 1) is that it will be anathema to those who hold that exact natural numbers are learnt, since this is tantamount to saying that exact natural number contents are the products of an innate and congenital ANS.

As an alternative, proponents of standard accounts can accept the existence of a congenital ANS but deny that it produces exact natural number concepts. For reasons we can now appreciate, this is the route that proponents of standard accounts 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 assert that, unlike exact number concepts, ANS representations "obscure the successor function" (Carey 2009: 295), and/or represent the number line as logarithmically compressed such that 5 is represented as less similar to 6 than 4 (DeWind et al. 2015). But now a different worry arises (Horn 2), for it is hard to see why conformity to the successor function would continue to constitute a natural way to interpret the number line from the perspective of a pre-numerate 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 or early emerging ANS. 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 it is assumed to emerge long before exact enumeration by virtually all theorists who have sought to specify how natural number concepts are learnt (Beck 2017; Carey 2009; Laurence & Margolis 2008; 2024; Shea 2011; Spelke 2017).[4] Likewise, we fail 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 nativism (Horn 1)[5], it remains unclear why

[4] 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.

[5] Barner (2017) appears to fall foul of Horn 1, despite his stated aim of avoiding this. He posits that ONE, TWO, and THREE are innate (565) and, furthermore, that children have an innate understanding that numbers are magnitudes and that magnitudes (often? always?) grow in a linear fashion. Thus, children have an innate understanding of the plus-one relation that holds between one, two, and three, and an innate understanding that this plus-one relation should be respected for all adjacent natural numbers when learning of these. To my ears, this collapses into Leslie et al.'s (2008) proposal according to which numeric conception relies on "an innately given recursive rule S(x) = x + ONE… also known as the successor function" (216) – a view Barner describes (unsympathetically) as an "extreme nativist" position (2017: 555).

conformity to the successor function will appear simple or natural to the child first learning to count (Horn 2). For instance, it would not help to follow Anobile et al. (2016) by 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 presents another deviant structuring of the number line for children to find simpler and more intuitive to adopt.

A more promising response would be to hold 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 rationally constrain wayward hypotheses about number and mathematics. But even bracketing recent evidence that pre-verbal infants (who are presumably not positioned to be taught how to count) already possess the competence to enumerate large numerical quantities exactly, 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 has motivated standard accounts of number learning.

For instance, we have 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 when asked. What is crucial to recognize, is that when children at these stages 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 (Carey & Barner 2019). So, much as Stromswold observes that children 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 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 matters 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 2019).[6]

The nativist can address both points. On (i), children's slow and incremental grasp of the words "one", "two", and "three" is seen to reflect difficulties mapping number words onto pre-existing number concepts (Spelke 2017; Margolis 2021); indeed, familiar biases in word learning predict difficulties of

[6] 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 is beyond the scope of this paper. Here, I 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 performance demands on participants (Izard et al. 2008).

precisely this sort (Clarke 2025). Moreover, an account of this sort still predicts that number words referring to small values will be learnt first, and in rank order, because the frequency with which a word referring to a number n is encountered follows a $P(n) \propto 1/n^2$ law, such that words for smaller numbers are encountered more 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 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 2019), researchers have found considerable variation in the knower-level that children reach before developing into cardinal-principle knowers. For instance, Krajsci and Fintor (2023) found that children regularly become four- and five-knowers, and occasionally six-, seven- or eight-knowers, prior to becoming cardinal-principle knowers (see also: Krajsci & Reynvoet 2024; Rouselle & Vossius 2021). Such variation is what we should expect if children learn to understand number words by incrementally mapping these onto an innately pre-structured sequence of natural number concepts, rather than the outputs of a parallel individuation system, with a strict set-size limit of three.

So, in sum: Humans acquire natural number concepts in specific ways, respecting the fact that natural numbers conform to the successor function. But standard accounts of number concept learning struggle to explain why. Following Chomsky's analysis of grammatical development, this is not simply because there are alternative interpretations of the number line that would be consistent with children's prior understanding and experience (*pace* Rips et al. 2006 & Rey 2014). It is because, otherwise, some of these alternative interpretations would strike pre-numerate children as simpler and more intuitive given their prior representational resources. Yet rather than being taught to avoid such deviant interpretations, extant evidence indicates that children never even consider these. Without some innately imbued appreciation for the natural numbers and their conformity to the successor function – e.g., an innate algorithm for generating arbitrarily large number representations in accord with this function (e.g., Leslie et al. 2008) – it seems hard to make sense of these results.

4. Is The New Riddle undermined by AI, or further supported by it?

The New Riddle of Concept Learning presents a challenge to standard accounts of numerical development according to which natural number concepts are learnt via rational processes. In Section 5, I will consider if and how The New Riddle applies to other conceptual domains. But, first, I want to consider an objection to the argument stemming from recent advances in AI.

This is pressing to consider, since it has been claimed that neural networks, like the transformer architectures modern large language models (LLMs) 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 abstracting from linguistic data on the internet. Similar concerns also afflict my argument as it pertains to the acquisition of number concepts. For while The New Riddle identifies an analogous poverty of the stimulus in the domain of number learning,

convolutional neural networks – without any pre-structured numerical competence – have been observed to abstract numerical contents.

To illustrate, Nasr et al. (2019) trained a convolutional neural network 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 importantly 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, the authors proceeded to test their 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, they presented the network with images containing 1-30 dots and, without further training, found that neurons in its feature-extraction network were already tuned to approximate numbers or 'numerosities'. Thus, after the network was trained to categorize images in its training data, individual neurons in the network were found to reliably fire in response to collections containing seven dots, or 22 dots, or 13 dots, etc. This was observed even though non-numerical properties of the displays (e.g., the cumulative area of the dots) were controlled for. This indicates 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 numerical representations were effectively learnt *tabula rasa*. 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.[7] For many contemporary empiricists, I suspect that it would be tempting to thereby conclude that number concepts can be abstracted through experience-driven learning.

This would, however, be a horrible mistake. The problem 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 to 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 deeper problem is that while the abovementioned networks achieve interesting levels of performance in the analysis of language and number respectively, Chomsky's poverty of the stimulus argument (and The New Riddle of Concept Learning which builds on this) does not simply challenge anti-nativists to explain patterns of

[7] 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.

success; it challenges critics to 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 who avail themselves of the abovementioned AI. There are two problems. 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 was children's failure to *ever* make errors of this sort which supported Chomsky's proposed need for an innate grammar. The same applies to the acquisition of number concepts: In Section 3, we saw that an account of numerical development must explain why children acquire natural number concepts that reflect conformity 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, with number-representing neurons' response profiles modelled by Gaussian curves on a logarithmic scale. So, while Nasr et al.'s network could seem to recommend some brand of numerical empiricism, it actually does quite the opposite: while *we*, numerate adults, might find it hard to imagine conceiving of the natural numbers as deviating from the successor function (perhaps, I'd argue, because of innate and hardwired conceptual machinery in this domain), Nasr et al. show that such deviant interpretations come naturally to an unconstrained learner, even when it lacks a pre-installed ANS that screams out in favor of these. Indeed, this last point underlines how deep The New Riddle runs. Were it not for results of this sort, one might suppose that The New Riddle could be avoided by simply invoking a slight touch of innate structure, which simply biases number learners away from mapping number concepts onto ANS representations, but otherwise leaves number learning unconstrained. Yet networks like Nasr et al.'s indicate that even this is unlikely to be enough; for even with a bias to avoid such a mapping, and a bias to disregard what the ANS dictates, Nasr et al.'s network demonstrates that a warped structuring of the number line might still come naturally to the learner.[8]

Might future neural networks fare better in this regard, avoiding these warped structurings, and acquiring exact number concepts that do 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 an innate tendency to construe numbers in this way was not baked in. Admittedly, we do not yet have a deep understanding of how such biases or knowledge gets 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 scientists and engineers tinker with the hyperparameters and architectures of neural networks is because their inbuilt structuring strongly determines *what* networks learn and constrain *how* they do so. If a network were to reliably acquire exact number concepts of a sort that humans employ in the math class, that systematically conform to the successor function, all

[8] This point is bolstered further by the widespread suggestion that an ANS has emerged in a great many creatures, including rats (Meck & Church 1983), pigeons (Riling & McDiarmid 1965), insects (Wittlinger et al. 2006), and fish (Agrillo & Bisazza 2018), not to mention non-human primates (Cantlon & Brannon 2006), while exact numeric conception has not. This is sometimes taken to show that the ANS is evolutionarily ancient and preserved across species (Carey 2009; Spelke 2022). However, it has been noted that the ANS now seems to be *so* widespread in nature that it is more likely to have emerged multiple times over, through a process of convergent evolution (Nieder 2021). Once again, this suggests that the emergence of numerical representations which respect the successor function is far from inevitable and may be significantly less likely to emerge than one which warps number space in the ways we have considered. That we find this surprising or counterintuitive is – to my mind – only further reason to suppose that our innate numerical machinery dictates that we conceive of natural numbers in the highly specific way we do.

while systematically avoiding the deviant interpretations adopted by Nasr et al.'s network and other systems, this would be compelling reason to conclude that its innate structuring somehow dictated this. In this respect, The New Riddle of Concept Learning seems to be largely insulated from advances in AI and cognitive development, in ways that make it quite unlike those more-familiar nativist arguments discussed in Section 1.

5. Scope of the Riddle

I have now argued that The New Riddle of Concept Learning supports the view that our grasp of exact natural number concepts is innately pre-structured, and I've proceeded to note that this conclusion is not undermined by recent advances in AI. But, when and where does The New Riddle apply? This is important to interrogate for two reasons. Firstly, 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. Secondly, it is important to clarify why The New Riddle does not overgeneralize, entailing the implausible conclusion that *all* concepts are innate products of our biology.

On the first of these points, consider how The New Riddle might be applied to theory of mind concepts. 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 general replicability (Lavelle 2022), The New Riddle confers non-trivial weight to a nativism in this domain, irrespective of how these matters pan out.

To give readers a taste for this: Consider that 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 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 some innate understanding of BELIEF and DESIRE, it is, however, hard to see how we might account for this. Non-mentalistic, behavioral rules can always be formulated to facilitate the prediction and explanation of observed actions (Povinelli & Vonk 2004). Thus, it is unclear why a creature who lacks mental state concepts would not find it simpler or more natural to interpret behavior in non-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 still indefinitely many models of the mental that 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 would strike us as natural and intuitive, if this were not innate. Indeed, this is particularly puzzling if our only innate model of the mental differs markedly from a mature grasp of mentality that must be learnt, as many theorists suppose (Apperly & Butterfill 2009; Butterfill & Apperly 2013; Burge 2018). Why? Because

this gives rise to a dilemma, analogous to that posed in Section 3 in the domain of number. For if such innate models of the mental dictate a stable understanding of belief-desire reasoning, akin to that found in adulthood, then the view collapses into full-blown BELIEF-DESIRE nativism (Horn 1); but if they do not, as is widely supposed, then it is unclear why humans would find it natural to conceive of BELIEF and DESIRE in the specific ways that they do, given that their only prior grasp of mentality is now seen to be screaming out in favor of a divergent model of the mental (Horn 2).

It might be replied that belief-desire reasoning is *taught* to children and, thus, passed down through generations (Heyes 2018). It is, however, hard to see how this could be, and for reasons that are rather more fundamental than in the case of number. The simple fact is: No one has ever managed to articulate the rules or principles that govern mature belief-desire reasoning (Maibom 2003). In fact, doing so seems 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 the abovementioned, 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). Once again, this suggests that our mature model of the mental is not one that comes simply or straightforwardly, absent innate structuring of the concepts in question.

Thus, I submit that a version of The New Riddle suggests that mental state concepts need to be innate, in a similar way to number concepts: Where natural number concepts, like SEVENTY-TWO, need to be the products of some innate, generative machinery, which ensures that numbers avoid conforming to a deviant arithmetic (e.g., Leslie et al. 2008), theory of mind concepts, like BELIEF, must also be innate, but in this case so as to ensure that their application to varied social predictions/explanations respects a basic logic that persists across time and place. In either case, innate structuring is needed to explain why these concepts are understood in uniform ways, despite the abundance of alternative construals or "canonical understandings" that might have emerged and would often seem to be simpler for a concept learner to adopt on non-nativist accounts.

We could go on. It is reasonable to speculate that The New Riddle might generalize to the acquisition of moral concepts whose application is stable across diverse populations (Mikhail 2002) and perhaps comparative concepts, like MORE and MOST (Clarke & Wellwood, forthcoming), which are deployed in opaque yet distinctive ways, under conditions where truth conditions are held constant (Knowlton et al. 2021). Nevertheless, the second point that I want to stress is that The New Riddle does not imply the innateness of *all* concepts. It thereby recommends a plausible brand of concept nativism, that should be taken seriously.

This is important to state, since the reader might now be wondering why The New Riddle would not dictate a view on which *all* concepts are innate. Suppose 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 unlearnt. At this point, you will have little sense of exactly what a carburetor is. But given general concerns about the inscrutability of reference, you might harbor a deeper worry: that further experience could never settle this. For instance, you might proceed to ask Google what a carburetor is and be told that it is a device that mixes air and fuel. But 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, coupled with a pre-existing capacity to appreciate this definition, it will be hard to see how there could avoid being a poverty of the stimulus, such that one's experience would always be consistent with myriad deviant interpretations of the concept (Rey 2014). Are we not, then, stuck concluding that CARBURETOR is innate in much the same way as SEVEN and BELIEF have been argued to be?

There is much to be said on this matter, but in short: We are not. In the case of natural numbers and theory of mind concepts, like BELIEF, I argued that the acquisition process is not or cannot be taught. But there is no reason why our concept of CARBURETOR could not be, with explicit instruction narrowing the space of possible interpretations. This might involve Googling of the above sort. It might also involve the testimony of an expert, charged with teaching car maintenance. Of course, for this guidance to help, there will still need to be innate structuring, which constrains how it gets interpreted (Gleitman 1990; Csibra & Gergely 2009); after all, rocks and other creatures are not positioned to learn about carburetors, no matter have much time and energy is poured into their education. Nevertheless, there is little reason to think that this innate structuring is in any way specific to the domain of carburetors or car parts; and when considered alongside one's instruction it is non-obvious why there remains a poverty of the stimulus to be overcome. This is because, I see no obvious reason to suppose that acquiring a mature (or even a fragmentary) conception of CARBURETOR involves subjects systematically and uniformly neglecting to consider simple or intuitive interpretations of what a carburetor is. So, while The New Riddle confers substantial weight to the conjecture that contested concepts, like SEVEN and BELIEF are innately pre-structured in nuanced and domain specific ways, the argument does not require that things spiral out of control, implying that *all* concepts are innately pre-structured in the way that some brands of concept nativism imply (e.g., Fodor 1980).

6. Conclusion

Arguments for concept nativism often turn on the idea that target concepts *could not* be acquired through learning, given an organism's situation, experience, or prior cognitive resources – that this would be *impossible*. However, in Section 1, I noted that a reliance on these assertions leaves nativists in a precarious position given recent advances in AI and developmental science, which purport to demonstrate that seemingly impossible methods of non-nativist acquisition might be achieved. For this reason, I suggested that concept nativists might be better served by The New Riddle of Concept Learning – developed here – at least in domains, like numeracy and theory of mind. Rather than trying to isolate an apparent incoherence or implausibility in the idea that these concepts might be learnt from experience, The New Riddle proceeds by asking how we might account for the highly specific and stable canonical understandings we acquire for these concepts, even when such understandings are not constrained by explicit teaching, and even when they do not constitute the simplest ways for learners to make sense of their experience (absent innate structuring in the relevant domain). The resulting argument builds upon Chomsky's insights regarding language, and prior arguments in the vicinity (Rips et al. 2008; Rey 2014), but – crucially – it does so in ways which avoid challenges facing these suggestions and in ways which avoid perceived problems for concept nativism that are posed by recent AI systems. For these reasons, my hope is that The New Riddle of Concept Learning might now take a more central role in rationalist accounts of the origin of (many) human concepts.

(Main text + footnotes: 10,707 words)

Works Cited:

1. Agrillo, C. & Bisazza, A. (2018). Understanding the origin of number sense: a review of fish studies. *Philosophical Transactions of the Royal Society, B Biol Sci.* 373(1740): 20160511.
2. Anderson, S.R. (1969) West Scandinavian vowel systems and the ordering of phonological rules (MIT dissertation).
3. 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.
4. 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.
5. Baillargeon, R. (1987). Object permanence in 3½- and 4½-month-old infants. *Developmental Psychology, 23*(5), 655–664.
6. Baillargeon, R., Scott, R. M., & He, Z. (2010). False-belief understanding in infants. *Trends in cognitive sciences*, *14*(3), 110–118.
7. Barner D. (2017). Language, procedures, and the non-perceptual origin of number word meanings. *Journal of child language*, *44*(3), 553–590.
8. Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"?. *Cognition*, *21*(1), 37–46.
9. Beck J. (2017). Can bootstrapping explain concept learning?. *Cognition*, *158*, 110–121.
10. Beck, J. & Clarke, S. (forthcoming). Cardinality, numerousness, and autoscaling: Revisiting the content and format of the approximate number system. In J. Park et al. (Eds.) *Numerical Cognition: Debates and Disputes*.
11. Berwick, R. C., Pietroski, P., Yankama, B., & Chomsky, N. (2011). Poverty of the stimulus revisited. *Cognitive science*, *35*(7), 1207–1242.
12. 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.
13. Boeckx, C. (2010). *Language in Cognition: Uncovering mental structures and the rules behind them.* Wiley Blackwell.
14. Brown, D., and Hippisley, A. (2012). *Network Morphology: A Defaults-based Theory of Word Structure*. Cambridge University Press.
15. Buckner, C. (2018) Empiricism without magic: transformational abstraction in deep convolutional neural networks. *Synthese* 195, 5339–5372.
16. Buckner, C. (2023). *From deep learning to rational machines: what the history of philosophy can teach us about the future of artificial intelligence*. New York, NY: Oxford University Press.
17. Burge, T. (2018). Do infants and nonhuman animals attribute mental states? *Psychological Review, 125*(3), 409–434.
18. Butterfill, Stephen A. (2017). Tracking and Representing Others' Mental States. In *Routledge Companion to the Philosophy of Animal Minds* (pp. 269--279). Routledge.
19. Butterfill, S.A. and Apperly, I.A. (2013), How to Construct a Minimal Theory of Mind. *Mind & Language*, 28: 606-637.
20. Butterworth, B., Reeve, R. et al. (2008). Numerical thought with and without words: Evidence from indigenous Australian children. *PNAS,* 105(35), 13179-84.
21. Cantlon, J. F., & Brannon, E. M. (2006). Shared System for Ordering Small and Large Numbers in Monkeys and Humans. *Psychological Science*, *17*(5), 401-406.
22. Carey, S. (2009). *The Origin of Concepts.* Oxford University Press.
23. Carey, S., & Barner, D. (2019). Ontogenetic Origins of Human Integer Representations. *Trends in Cognitive Sciences.* 23(10):823-35.
24. Chomsky, N. (1965). *Aspects of the theory of syntax*. MIT Press, Cambridge, MA.
25. Chomsky, Noam A. (1980). *Rules and representations*. Oxford: Blackwell.
26. Clarke, S. (2023). Compositionality and constituent structure in the analogue mind. *Philosophical Perspectives*, 37: 90–118.
27. Clarke, S. (2025). Number nativism. *Philosophy and Phenomenological Research*, 110, 226–252.

28. Clarke, S., & Beck, J. (2021). The number sense represents (rational) numbers. *Behavioral and brain sciences*, *44*, e178.
29. Clarke, S. & Wellwood, A. (forthcoming). Three brands of concept nativism. *Analysis*.
30. Clarke, S. & Wellwood, A. (MS.). One TWO or two?
31. Csibra, G., & Gergely, G. (2009). Natural pedagogy. *Trends in cognitive sciences*, *13*(4), 148–153.
32. Dehaene, S. (2011). *The Number Sense: How the Mind Creates Mathematics: How the Mind Creates Mathematics*. Oxford University Press.
33. Dehaene, S., & Mehler, J. (1992). Cross-linguistic regularities in the frequency of number words. *Cognition*, *43*(1), 1–29.
34. 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.
35. 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.
36. Feigenson, L., Dehaene, S., & Spelke, E. (2004) Core systems of number. *Trends in Cognitive Science* 8(7):307-314.
37. Fodor, J. (1975) *The Language of Thought*, New York: Thomas Y. Crowell.
38. Fodor, J. (1980) The present status of the innateness controversy. In *Representations: philosophical essays on the foundations of cognitive science*. Cambridge: MIT Press. pp. 257-316.
39. Fodor, J. (1994) *The Elm and the Expert*, Cambridge, MA: MIT Press.
40. Fodor, J. (1997). Special Sciences: Still Autonomous After All These Years. *Philosophical Perspectives*, *11*, 149–163.
41. Fodor, J. (2008). *LOT 2: The Language of Thought Revisited*. Oxford: Oxford University Press.
42. Fodor, J. (2010). Woof, Woof. Review of *The Origin of Concepts* by Susan Carey, *The Times Literary Supplement*, October 8: pp. 7–8.
43. Gleitman, L. (1990). The structural sources of verb meanings. *Language Acquisition,* 1(1), 3-55.
44. Gordon, P. (2004). Numerical Cognition Without Words: Evidence from Amazonia. *Science*, 306(5695), 496–499.
45. Gunasekar, S., Zhang, Y., Aneja, J., Mendes, C. C. T., Del Giorno, A., Gopi, S., Javaheripi, M., Kauffmann, P., de Rosa, G., Saarikivi, O., Salim, A., Shah, S., Behl, H. S., Wang, X., Bubeck, S., Eldan, R., Kalai, A. T., Lee, Y. T., & Li, Y. (2023). *Textbooks are all you need*. arXiv. https://arxiv.org/abs/2306.11644
46. Halle, M. (1997). Distributed morphology: Impoverishment and fission. *MIT working papers in linguistics.* 30:425-49.
47. Heyes C. (2014). False belief in infancy: a fresh look. *Developmental science*, *17*(5), 647–659.
48. Heyes, C. (2018). *Cognitive gadgets: The cultural evolution of thinking*. Harvard University Press.
49. Izard, V., Pica, P., Spelke, E. & Dehaene, S. (2008) Exact Equality and Successor Function: Two Key Concepts on the Path towards Understanding Exact Numbers, *Philosophical Psychology*, 21:4, 491-505.
50. 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.
51. 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.
52. Kodner, J., Payne, S., & Heinz, J. (2023). *Why linguistics will thrive in the 21st century: A reply to Piantadosi,* arXiv: 2308.03228.
53. Krajcsi, A., & Fintor, E. (2023). A refined description of initial symbolic number acquisition. *Cognitive Development*, *65*, 101288.
54. Krajcsi, A., & Reynvoet, B. (2024). Miscategorized subset-knowers: Five- and six-knowers can compare only the numbers they know. *Developmental science*, *27*(1), e13430.
55. Laurence, Stephen & Margolis, Eric (2001). The poverty of the stimulus argument. *British Journal for the Philosophy of Science* 52 (2):217-276.
56. Laurence, S., & Margolis, E. (2002). Radical concept nativism. *Cognition*, *86*(1), 25–55.
57. Margolis, E., & Laurence, S. (2008). How to learn the natural numbers: inductive inference and the acquisition of number concepts. *Cognition*, *106*(2), 924–939.
58. Margolis, E., & Laurence, S. (2024). *The Building Blocks of Thought: A Rationalist Account of the Origin of Concepts*. Oxford: Oxford University Press.

59. 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.
60. 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.
61. Leslie, A. M., Gelman, R., & Gallistel, C. (2008). The generative basis of natural number concepts. *Trends in Cognitive Sciences*, 12, 213–218.
62. Lewis, D. (1994). Reduction of Mind. In S. Guttenplan (ed.), *A Companion to Philosophy of Mind*, Oxford: Blackwell, pp. 412–31.
63. Libertus ME, & Brannon EM. (2010). Stable individual differences in number discrimination in infancy. *Developmental Science*. 13(6):900-906.
64. Lipton, J. S., & Spelke, E. S. (2003). Origins of number sense. Large-number discrimination in human infants. *Psychological science*, *14*(5), 396–401.
65. 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.
66. Maibom, H. (2003), The Mindreader and the Scientist. *Mind & Language*, 18: 296-315.
67. 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.
68. Margolis, E. (2021). The Small Number System. *Philosophy of Science*, 87(1), 113–134.
69. McCrink, K., & Wynn, K. (2004). Large-number addition and subtraction by 9-month-old infants. *Psychological science*, *15*(11), 776–781.
70. Meck, W. H., & Church, R. M. (1983). A mode control model of counting and timing processes. *Journal of experimental psychology. Animal behavior processes*, *9*(3), 320–334.
71. Mikhail, J. (2007). Universal moral grammar: Theory, evidence and the future. *Trends in Cognitive Sciences, 11*(4), 143–152.
72. Munakata, Y., McClelland, J. L., Johnson, M. H., & Siegler, R. S. (1997). Rethinking infant knowledge: Toward an adaptive process account of successes and failures in object permanence tasks. *Psychological Review, 104*(4), 686–713.
73. 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.
74. Nieder A. (2021). The Evolutionary History of Brains for Numbers. *Trends in cognitive sciences*, *25*(7), 608–621.
75. Pater, J. (2019). Generative linguistics and neural networks at 60: Foundation, friction, and fusion. *Language*, 95(1):e41–e74.
76. Peacocke, C. (1992). *A Study of Concepts*. MIT Press.
77. 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.
78. 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
79. 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.
80. Pinker, S. (1984). *Language learnability and language development*. Harvard University Press.
81. Pinker, S. (1994). *The language instinct*. New York: William Morrow.
82. Povinelli, D.J. and Vonk, J. (2004). We Don't Need a Microscope to Explore the Chimpanzee's Mind. *Mind & Language*, 19: 1-28.
83. Prinz, J. (2002). *Furnishing the Mind: Concepts and Their Perceptual Basis*. MIT Press.

84. Qu, C., Szkudlarek, E., & Brannon, E. M. (2021). Approximate multiplication in young children prior to multiplication instruction. *Journal of Experimental Child Psychology, 207,* 105116.
85. Rawski, J. and Heinz, J. (2019). No Free Lunch in Linguistics or Machine Learning: Response to Pater. *Language*, 95(1):e125–e135.
86. 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.
87. Rilling, M. & McDiarmid, C. (1965). 'Signal Detection in Fixed Ratio Schedules'. *Science, 148, 526–7.*
88. Rogers, T.T., & McClelland, J.L (2004). *Semantic Cognition: A Parallel Distributed Processing Approach*. Cambridge, MA: MIT Press
89. Rousselle, L., & Vossius, L. (2021). Acquiring the Cardinal Knowledge of Number Words: A Conceptual Replication. *Journal of Numerical Cognition*, *7*(3), Article e7029.
90. 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.
91. Sachdeva, N., Coleman, B., Kang, W., Ni, J., Hong, L., Chi, E., Caverlee, J., McAuley, J. & Cheng, D. (2024). How to Train Data-Efficient LLMs. arXiv, https://arxiv.org/abs/2402.09668
92. Samuels, R. & Snyder, E. (2024). *Number Concepts: An Interdisciplinary Inquiry*. Cambridge University Press.
93. Schneider, R. M., Pankonin, A., Schachner, A., & Barner, D. (2021). Starting small: Exploring the origins of successor function knowledge. *Developmental Science, 24*(4), Article e13091.
94. 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.
95. Smith, E.E. and Osherson, D.N. (1984), Conceptual Combination with Prototype Concepts. *Cognitive Science*, 8: 337-361.
96. Spelke, E. S. (2017). Core knowledge, language, and number. *Language Learning and Development, 13*(2), 147–170.
97. Spelke, E.S. (2022). *What Babies Know: Core Knowledge and Composition Volume 1*. Oxford University Press.
98. 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
99. van Dantzig, S., Raffone, A., & Hommel, B. (2011). Acquiring contextualized concepts: a connectionist approach. *Cognitive science*, *35*(6), 1162–1189.
100. Warstadt, A., Singh, A., and Bowman, S. R. (2019). Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.
101. 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.
102. Wittlinger, M., Wehner, R., & Wolf, H. (2006). The ant odometer: stepping on stilts and stumps. *Science (New York, N.Y.)*, *312*(5782), 1965–1967.
103. Wynn, K. (1990). Children's understanding of counting. *Cognition, 36*(2), 155–193.
104. Wynn, K. (1992). Addition and subtraction by human infants. *Nature* 358, 749–750.
105. Xu, F., & Spelke, E. S. (2000). Large number discrimination in 6-month-old infants. *Cognition*, *74*(1), B1–B11.
106. Yang, C. (2016). *The price of linguistic productivity*. Cambridge, MA: The MIT Press.
107. 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.