A Parallel-Distributed Processing Approach to Mathematical Cognition[1]

James L. McClelland, Kevin Mickey, Steven Hansen, Arianna Yuan and Qihong Lu

Stanford University
February 18, 2016

How should we think about the nature of our knowledge of mathematical concepts, and the mechanisms we use to learn and use these concepts when we do mathematics? Here we describe a perspective on the answers to these questions and a future research program to address them that is grounded, in part, in the parallel distributed processing [PDP] approach to cognition and learning [1,2] implemented in artificial neural networks. We begin with a more basic question, namely the nature of mathematics and mathematical reasoning, and proceed from there to consider mathematical knowledge, learning, and thinking, stressing the roles of culture and experience in the creation and learning of mathematics. We then review the PDP perspective on the nature of knowledge and learning, and consider how it can address many findings in the mathematical cognition literature, and how it provides alternative ways of understanding what nature may provide and what nurture may create. Next we discuss the exciting challenges facing the approach and how they might be addressed, organizing the discussion around the question: How might a neural network-based approach meet the challenge of learning to achieve a level of competence in algebra and geometry sufficient to pass a high-school proficiency exam in geometry, in a human-like way, from experiences similar to those of human learners? We conclude with a discussion of the implications of the approach for learners and teachers of mathematics and for the processes of teaching and learning.

What is Mathematics and what is Mathematical Thinking?

One widely held view of the nature of mathematics is that it is essentially formal. This view is historically associated with Bertrand Russell [3], who famously said 'all mathematics is symbolic logic'. This view may have contributed to the enthusiasm, in the early days of cognitive science, for a formal perspective on the nature of thought in general. Fodor & Pylyshyn [4] captured these ideas by claiming that all systematic cognition is symbol processing defined as the manipulation of structured expressions according to structure-sensitive rules. On this view, mathematics is essentially thought itself, and thought mathematics. By the early 1970's, computer programs were written that could process any solvable integro-differential equation using this approach, producing mathematical results by obeying a system of structure-sensitive rules without regard to the meanings of the expressions they manipulated [5]. These developments may have served to reinforce the view that mathematics is, inherently, a matter of structured expressions and structure-sensitive rules, and the related view that mathematical thought is a matter of manipulating such expressions according to such rules.

[1] Based on Heineken Prize lectures given by JLM at the University of Amsterdam, Autumn, 2014, and at the Cognitive Science Society Meeting in Pasedena CA, August, 2015. We thank David Landy, members of the PDP Research Group at Stanford, and many others for useful discussions.

The essentially formalist perspective contrasts with an alternative perspective that is adopted by many mathematicians, however. The mathematician Tristan Needham [6] likens studying mathematics as symbol manipulation to studying music without ever hearing a note. To him, the things mathematical expressions *refer to* – things we can often draw on paper and see in our minds eye – are the real objects of mathematical thought. In line with this, Roger Shephard (a mathematical thinker whose intuitions shaped several branches of cognitive science) has claimed that the mental manipulation of visualized objects of thought could "convince us of the validity of mathematical and physical laws" [7] such as the Pythagorean Theorem (Figure 1). On this view, mental operations performed on visualized objects of mathematical thought are more central to mathematical reasoning than is symbol manipulation [8], and play a key role in discovering and proving theorems in mathematics. The kinds of transformations Shephard stresses (mental translation and rotation [9]) are shape- and area-preserving when applied to real objects, and if these properties are maintained when these transformations are carried out in the mind, they allow mental transformations to reveal the truth of relationships such as the one expressed by the Pythagorean theorem. This view is reflected in systems of geometric and mathematical reasoning including *Transformational Geometry* [10] and in approaches to teaching mathematics that stress the importance of viewing mathematical expressions as statements about *quantities* [11], defined as measurable properties of the objects of mathematical thought, such as the area of a square or the cardinality of a set. Mental operations on visualized objects then allow intuition and insight to inform mathematical reasoning and may be the basis for many discoveries in mathematics, even if formal proofs are often presented in terms of long symbolic derivations that obscure the underlying intuitions [12]. Here, obviously, mathematical thought – and thought in general – are far more than merely symbol manipulation.

But if the objects of mathematical thought are not symbols, what, exactly, are they? According to the Platonic view, these objects exist as ideals independent of human minds. The Platonic view captures the ideal nature of these objects – the fact that, for example, no drawn or constructed circle actually conforms perfectly to the ideal circle as defined in mathematics. The Platonic view does not, however, help us understand why even the simplest aspects of mathematics emerged late in human evolution and are not universal across cultures. An alternative perspective is provided by the mathematician Rubin Hirsh [13], himself building on ideas of others [14]. On this view, mathematics, and the objects found in mathematics, are the products of human cultures. Mathematics consists of a set of constructed reasoning systems that arise initially for utilitarian purposes when cultures advance to the point where they need them and that then gradually evolve and become systematized over time. The needs that give rise to mathematical systems may include record keeping or reasoning about culturally relevant properties of objects or collections of objects, such as the surface area of a flat object (how big is a piece of land) or the cardinality of a set of objects (how many baskets of grain are you storing for me?). The systems rely on facts about these properties, such as the fact that the shape of an object is conserved under translation, rotation and reflection, and the fact that the cardinality of a set of objects is conserved under rearrangement of its members. Systems that allow reasoning about these properties come into use prior to their complete formal characterization, do not depend on a fully adequate formal characterization to be of use, and are progressively refined and formalized over time [14]. Of course, cultures of mathematics and mathematics education grow up around these systems, and authorities and

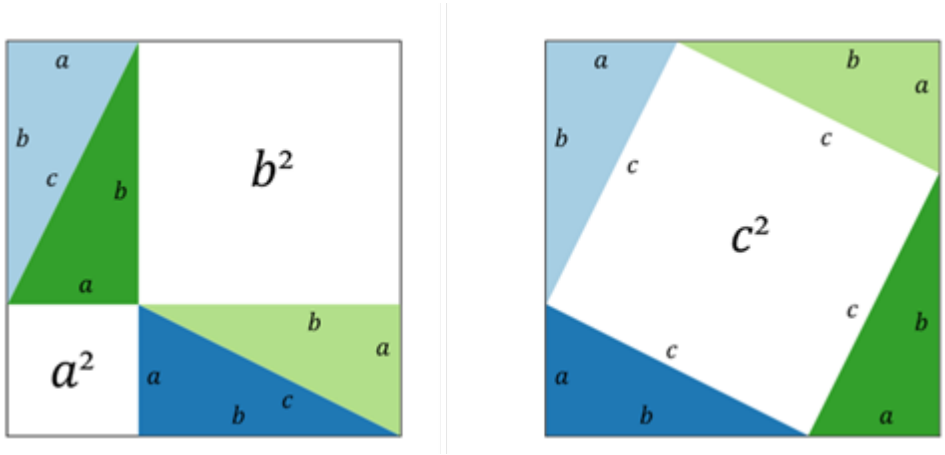


Figure 1. A visuospatial proof of the Pythagorean Theorem similar to the one in [7]. This form of the proof may have pre-dated Pythagoras. The theorem states that the sum of the areas of the squares on the perpendicular sides (a and b) of a right triangle is equal to the area of the square on the hypotenuse (c) of the triangle. The drawing on the left shows four identical triangles arranged within an enclosing square, leaving regions of area a^2 and b^2 uncovered. After moving some of the triangles, the uncovered region takes the form of a square on the hypotenuse of area c^2. The proof depends on the fact that area is conserved under translation, a fact that may serve as an intuitive basis for understanding the Pythagorean Theorem.

established practices within these cultures then constitute additional important constraints on what mathematics is in practice.

We take Hirsh's characterization of the nature of mathematics as the starting point of our own efforts to understand human mathematical cognition. By starting with the view that mathematics consists of culturally-constructed reasoning systems that arise for initially utilitarian purposes, we accept a role for formalism in mathematics as well as roles for symbol manipulation and visuospatial intuition in mathematical thinking. What we suggest, however, is that both symbol manipulation and visuospatial reasoning may depend as much on acquired abilities humans develop in response to the demands of culture and education as on inherent characteristics of the human mind. Our approach reflects the view that the human ability to exploit the culturally constructed reasoning systems that constitute mathematics need not be an inherent consequence of previously evolved features of the human mind that are specifically attuned to the demands imposed on them by mathematics *per se*, even though this ability draws on both specific as well as general characteristics of diverse processing and learning systems in the brain.

Origins of Mathematical Knowledge

Just as there are several views about the essential nature of mathematics, there are also several views about the origins of mathematical knowledge. Within the Platonic philosophical tradition, the idealized objects of mathematical thought are innately known. This perspective is presented in Plato's *Meno* [15], in which Socrates purportedly demonstrates that a completely uneducated slave possesses a pre-

existing understanding of the concept of area and of how to construct a square with twice the area of a given square, even though the slave demonstrates several misconceptions during the Socratic dialog. Just as the Platonic view is unhelpful in allowing us to understand why mathematics emerged so late in human evolution, it also fails to help us understand why many aspects of mathematics are very hard for most people to learn. Indeed, a recent study suggests that the slave's misconceptions are common and those who have them fail to gain a productive understanding they can apply to carry out the construction of a square with twice the area of a given square on their own [16]. As we will discuss at more length below, misconceptions and pervasive patterns of error are found at all stages of mathematics learning. A similar challenge arises for a symbolic approach to mathematical cognition: If the mind is inherently a symbol processing machine and if mathematics is symbol processing, then the mind should be well-prepared for mathematics, and mathematics should not be so difficult for so many people to learn.

Among contemporary researchers who consider the origins of mathematical knowledge, many still stress the role of innate constraints that are *a priori* and *universal* (as proposed in Kant's *Critique of Pure Reason*[17]) and that serve as the foundation for mathematical learning. However, there is a range of views about what these innate constraints may actually be [18], [19], [20], [21]. We are certainly open to the possibility that human mathematical reasoning systems may build on intuitions supported in part by innate constraints. Even more important, in our view, is the effort to understand the role experience may play in shaping mathematical ability and mathematical intuition. This perspective is consistent with a broad tradition on the role of culture in cognition more generally [22,23]. The experiences to which we refer include experiences consistent with very basic invariances such as the continued existence of objects that have gone out of view, the conservation of number under spatial rearrangement, and the conservation of shape of rigid objects under translation, rotation, and change of viewpoint. These invariances are properties of objects and sets of objects in the world. Culture (including institutionalized educational practice) shapes exposure to them and can create structured task settings that isolate, idealize, describe, and quantify them. Systems for counting exemplify this idea. Counting is not culturally universal [24], and there is evidence that adults from cultures lacking exact number systems and children in cultures that do have such systems have at best an incomplete understanding of what it means for two sets of items to have exactly the same number of members [25],[26], consistent with the view that the very idea of exact number is acquired. Furthermore, counting systems vary widely in details across cultures, and vary extensively in how well they support systematic numerical computations [27]. The place-value system now in use in the internationally shared culture of science, mathematics, and computation and in the educational systems that prepare learners for a place within this culture is quite a recent development [13].

Origins of learning in joint activities structured by cultural practices. A key element of our perspective is the idea that cultural systems for mathematical reasoning can be grounded in operations applied to real or depicted objects or sets of objects that can be part of the experience of a learner. The situations in which these experiences arise communicate cultural practices through interactions between a learner and a more experienced member of the child's culture, such as a caregiver or a teacher [28]. Consider a situation in which a caregiver is showing a child a picture of three frogs. The caregiver can say 'look,

three frogs – let's count them – 1,2,3', touching each frog in turn as she counts. The caregiver can also ask the child to count the frogs, observing the pointing gestures the child makes as well as the final result of the count. In this way the caregiver can provide the child with valid examples of counting events and determine whether the child adheres to valid principles of counting at each step in the counting process [29]. As a second example, early arithmetic is often taught using fingers or tokens such as beans or coins [30]. A child may learn to add 3 plus 2 by first counting out three fingers on one hand, then two on the other, then counting all of the fingers raised. Beyond simple counting and addition of small numbers, further examples of externalized representations of number are the abacus and the standard place-value system for representing number that all students now learn in school. Another example is the system for manipulation of algebraic expressions taught starting in middle school. All of these are taught, and learned, through experiences in which teachers and learners can watch each other as they illustrate and attempt to replicate valid operations on externalized representations.

Similar points can be made about other mathematical systems for quantifying and for representing continuous amount (using a ruler or a scale to measure the length or weight of an object) and for reasoning about geometric objects and their properties. Many babies in conventionally educated families may gain relevant experience through the manipulation of toys, such as a shape-sorter, in which the child can begin to gain experience about the properties of idealized shapes used in geometry. Euclidean geometry is grounded in part on construction of figures relying on a straightedge and compass. Thus, the visuospatial manipulation of the mentally represented objects of thought discussed above may have much of its origin in culturally constructed systems for the manipulation and construction of such objects and diagrams of such objects. In general, we suggest that internalized mathematical cognition may originate largely from experiences interacting with real or depicted objects starting with informal early experiences and progressing through structured experiences provided in educational materials and settings. We can *learn* to represent and manipulate objects of mathematical thought in our minds, building from experience manipulating real objects and externalized depictions of them.

The contrast between the different perspectives on the basis of mathematical cognition is more a matter of emphasis – virtually no one would hold today that nature and nurture are not both essential to the emergence of mathematical ability. Accepting this, we still see reason to stress the importance of experience. This message is brought out in considering different conceptions of the concept of a 'mental number line' [31–33]. One very prominent view holds that the mental number line is an innate structure localizable to a region of the parietal lobe that intrinsically links magnitude to position to space [34,35]. An alternative view holds that the mapping of number onto space is a late cultural invention (first attested in the 17th century) [36], one that arises gradually in development through structured experiences with numbers and with cultural practices for mapping of numbers onto space [31,33] and depends for its use on an acquired understanding of the relationships between numbers [31]. While the innate and acquired number-line views are not necessarily mutually exclusive, a consideration of a role for experience appears necessary to address a wide range of recent findings showing how performance marking numbers on a line gradually changes with development and number knowledge [37–39]. Among other things, young children may fail to place numbers accurately on a line from 0 to 100

because they lack a clear conception of the meaning of 100 and thus cannot gauge how far numbers like 90 should be placed from an anchor point at 100, or because they rely less on landmarks e.g., at 50 on a line from 0 to 100 [38]. Greater appreciation of the meaning of 100 and of its relation to other numbers may thus support increases in line marking accuracy. Viewed in this light, individual differences in the precision of number line marking may be a reflection, rather than a cause, of increased understanding of numbers and of experience mapping number onto space. Indeed such experience appears to influence number-line marking [40].

The view that so much of mathematical cognition is a consequence of learning from experiences that vary from culture to culture (or from individual to individual within a culture) may fly in the face of the intuitions of expert mathematicians, who attest to the immediacy of a sense of understanding that can arise in the tutored mind when watching, for example, an animated version of the proof Shepard presented of the Pythagorean theorem. This immediacy may provide a phenomenological basis for the view that aspects of mathematical understanding are immediate or even innate. According to the current perspective, this immediacy reflects a consequence of what the philosopher of science Howard Margolis has called *acquired habits of mind* [41], which allow conventionalized diagrams or even symbolic mathematical expressions to come to be mapped automatically (without effort or intention) through repeated practice to mental representations structured by the internalization of culturally constructed systems of thought. These experiences can create the illusion that a mathematical truth is known directly even though its appreciation depends crucially on an extensive body of accumulated experience.

The Nature of Knowledge, Its Origins, and how it is Affected by Learning

If so much of mathematics is acquired, how should we think about the nature of what we learn and the process by which we learn it? A classical view is that knowledge consists of explicit systems of rules or propositions – representations that we appeal to as such as we think [42,43]. Developmental psychologists have often characterized initial knowledge as structured ensembles of propositional statements, and subsequent development is viewed as a process of enriching or possibly re-structuring such representations [44,45]. Furthermore, given that knowledge is represented in the form of rules or propositions, there must be a moment of rule induction – what Pinker [46] called a 'Eureaka moment' – in which the rule is 'discovered'. This view still appears to be held by many leading researchers in the field of mathematical cognition. For example, in a recent discussion of children's acquisition of the cardinality principle (CP, see below), the authors wrote [25]:

> *The exact nature of the insight that children experience when they reach the state of CP-knower is unclear. In order to understand it better, we need to determine what children know just before this insight, what triggers it, and what they finally derive from their newly acquired numeric competences.*

We argue for an alternative to this perspective that arises from work within the parallel distributed processing (PDP) framework [1,2]. According to this framework, our cognitive abilities – including mathematical abilities – depend on processes that arise through the propagation of activation among

interconnected networks of neurons. PDP models do not attempt to capture cognition at a detailed neurophysiological level – instead, they focus on a functional-level description, aimed at modeling the time course and outcome of cognition and at capturing how initial conditions and experience my shape cognition, using networks typically containing far fewer simulated neuron-like processing units than are actually used in the human brain. The mechanisms that govern learning in such models are also characterized at a functional level, and rely on assumptions about the propagation of learning signals whose biological realization is not yet fully understood [47]. This framework allows us to rethink the very nature of knowledge itself, and then from this to re-think the starting place for knowledge acquisition, and the process by which knowledge is acquired and cognitive abilities change.

The knowledge is in the connections. According to the PDP approach, the knowledge that governs processing is not to be found in explicit rules but in the pattern of connections among the neurons that participate in our cognitive activities. Such connections, on this view, underlie perceptual processes such as mapping sensory inputs onto more abstract representations as well as conceptual and linguistic processes such as mapping linguistic or perceptual input onto underlying representations of objects or events and the generation of language outputs such as the mapping from spelling or meaning to sound. These models do not actually employ rules as such, although they can often be *described as though* they behave in accordance with rules.

Initial conditions. If knowledge is in connections, then the initial state of a processing system is to be found in the initial pattern of connections. A key contribution to the development of a connectionist perspective on this matter was provided in *Rethinking Innateness* [48]. Here it was noted that the initial conditions of a neural network could dramatically influence the time-course and outcome of learning, and could even affect performance of neonates in a range of tasks, without encoding any initial knowledge in propositional form. Exactly how initial tendencies are captured in connections can then be explored through neural network modeling. It is possible for a neural network modeler to build specific assumptions about how content is processed directly into the wiring of a neural network, but it is also possible to explore how characteristics of processing that are revealed in experiments might arise from generic network properties or from variation in meta-parameters of different parts of a network. An early application of this idea was the demonstration that subcortical center-surround cells and, at deeper layers, neocortical edge detectors might arise naturally, driven only by random spontaneous activity in the retina (occurring in an unborn fetus), a locality constraint on connectivity and Hebbian synaptic plasticity, if deeper layers enforce a greater sparsity of representation than more superficial layers, through a meta-parameter that regulates the overall firing rate of neurons in the layer [49].

Gradual learning. Central to our own effort is the idea that much of knowledge acquisition is to be understood as resulting from gradual connection strengthening processes in relatively generic neural networks. The PDP model of the English past tense [50] provided an early instantiation of this view, in that the model acquired the ability to behave in approximate accordance with explicit rules of English past-tense formation without recourse to an explicit representation of the rule. Though the first version of this model had limitations [51], subsequent models overcame them, and other models applied the same ideas to learning the mapping from the spelling of a word to its pronunciation [52,53]. Another example is the neural-network based TD-gammon model [54] that learned to exceed human and other

machines' backgammon abilities through the use of reinforcement learning to gradually adjust strength of connections in a neural network trained only with feedback on the eventual success or failure of its play. Recent deep neural networks that learn to classify objects [55], translate one language into another [56], and to exceed human abilities in many computer-based action games [57] are more modern instantiations of these same ideas. Interestingly, the deep neural networks that master these action games exhibit patterns of play that can be characterized at a high level as discovered *strategies,* even though no explicit strategic level thinking is occurring in the neural network, in the same way that the past-tense model did not really infer an explicit linguistic rule.

We suggest that the learning characteristics of PDP models may be helpful for understanding how humans acquire mathematical abilities for three reasons: They exhibit initially gradual learning that can accelerate accounting for changes in readiness to learn; gradually increasing strength, robustness, and automaticity; and sensitivity to statistical biases in their experiences. While inter-related, these three characteristics of neural networks are worth considering separately, since each can help explain distinct aspects of mathematical cognition.

Changes in readiness to learn. Neural network models generally start to learn slowly, and acquire structured knowledge over thousands to millions of relevant experiences, while also showing a strong non-linear dependency between the strength of existing knowledge and how rapidly they learn. This can help us understand why learning can seem very slow at times, why and how rapid-insight-like transitions can occur, and how the ability to learn can depend on what we already know, addressing Vygotsky's ideas that learning depends on presenting learners with experiences within their 'zone of proximal development' [23].

These properties are all illustrated by a simple neural [58] model of learning about the roles of weight and distance in determining which side of a scale would go down if it were free to move (Figure 2). While Siegler [59] initially characterized children's knowledge of the balance scale according to a system of rules, several features of the behavioral data from children's and adults' performance with the balance scale suggested that the underlying knowledge might be graded and that its acquisition is more gradual than expected under a rule-based characterization [60]: (i) though most children could be assigned to one of Siegler's rules, the pattern of their responses deviated from these rules in subtle ways; (ii) children show graded sensitivity to degrees of difference in weight or distance not adequately captured by the rules and (iii) studies allowing a continuous rather than categorical response establish that the sensitivity to both cues gradually strengthens with age. Most relevant here was a study [59] examining the transition from Siegler's Rule 1 (if the weights differ, the side with the greater weight goes down; otherwise the scale will balance) to his Rule 2 (the first part of Rule 1 is retained, but if the weights are equal, distance is considered). Older Rule-1 children were able to progress from a series of experiences that challenged their pattern of responding while younger Rule-1 children were not. The

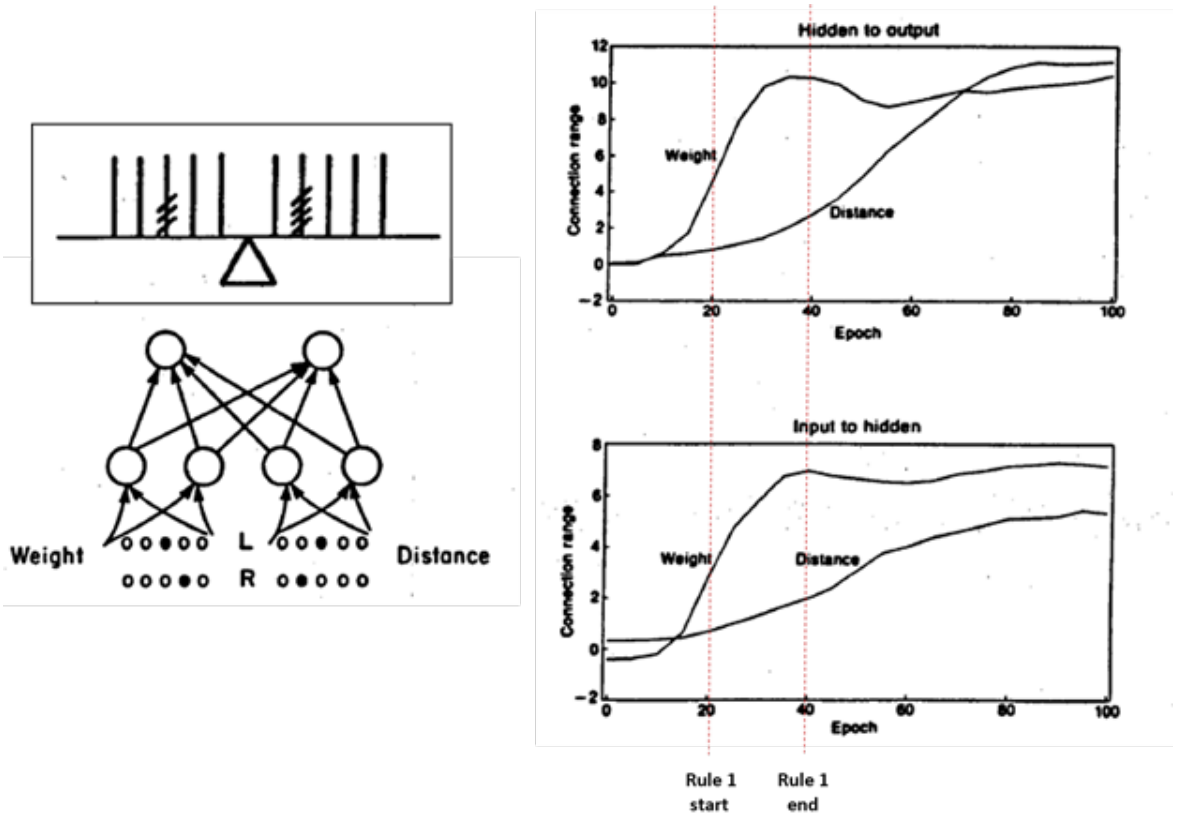


Figure 2. Top Left, an example balance scale problem of the type used in [59] and other studies. Bottom left, the neural network model of balance scale learning [58,60]. At right, the gradual development in the model of connection-based knowledge about the roles of weight and distance in balance is illustrated separately for the hidden-to-output connection weights and the input-to-hidden connection weights. The quantity plotted on the vertical axis is a measure of the extent to which the connection weights in the network vary with variations in weight or distance.

simple neural network model shown in Figure 2 captured all aspects of these findings, and its characterization of the developmental process has held up [61] against subsequent empirical evidence that appeared to some [62] to be inconsistent with it. Importantly, change in the connection weights is essentially continuous and is very slow at first but accelerates as knowledge builds up, leading to apparently abrupt changes that still show signs of graded cue sensitivity. This is why the model, like children, progresses from experiences that challenge a 'Rule 1' understanding toward the end of its Rule 1 phase, but fails to do so when challenged earlier in learning. A rigorous mathematical analysis of why this kind of pattern occurs in deep neural networks is available [63].

Gradually increasing strength, robustness, and automaticity. In a neural network model, strength of knowledge and learning is always a matter of degree. Once knowledge in the system is strong enough to generate an appropriate response, it nevertheless continues to strengthen, gradually becoming more robust and more automatic. We view the acquired speed and relative automaticity of processing that can occur in neural network models as providing the basis for the gradual entrenchment of processing

habits that effectively automatize certain basic aspects of number processing, such as the processing of the magnitude of a numerical stimulus. This is reflected in the numerical Stroop effect [64], the tendency for numerical value to interfere with processing the physical size of a numeral, including the emergence of this effect over development [65]. A similar tendency may also be reflected in the SNARC effect [66], where association of numbers to space and also to fingers used to count smaller vs larger numbers [67] may become entrenched sufficiently that they influence processing even when magnitude and numerical value are irrelevant. In a neural network, automaticity, like strength itself, is a matter of degree, and is reflected in (a) how quickly activation occurs, (b) how much the speed and outcome of activation depends on top-down control, and (c) how strongly the process will interfere with other processes [68]. Such features appear to characterize the effects of extensive practice on human perception and cognition, and have been used to account for the role of experience in the Stroop interference effect [69]: that is, the finding that a process that was itself initially susceptible to interference and did not initially interfere with another process could come, through practice, to exhibit both reduced susceptibility to interference and an increased tendency to interfere with another process. More speculatively, we suggest that practice leads to entrenched habits of mind [41] that can gradually result in an acquired tendency to 'just see' that certain mathematical relationships hold without apparent intervening processing stages.

Sensitivity to statistics of experience. Deep neural networks show sensitivity to statistical properties of experience, such as the frequency and typicality of the items they encounter as they learn [50]. These effects are very striking during early learning – frequent and typical patterns of responding are the first to be acquired, and can initially over-ride correct responses to infrequent or atypical items. For example, a deep neural network model of natural kind semantics [70] exhibits strong overgeneralization errors, attributing eyes and legs to animals that do not have them and labeling a less familiar animal (e.g. *goat*) with a more familiar animal's name (*dog*). Similar tendencies are seen in young children [71,72]. But as learning progresses, the networks eventually learn to respond correctly (albeit less reliably and robustly) even to these infrequent or atypical items, thus giving rise to U-shaped developmental trajectories similar to patterns seen in human data [73] in several domains. As discussed below, many findings in mathematics learning are well captured by these aspects of neural networks.

In summary, the ability to act in a way that is consistent with principles or rules can emerge through a sub-symbolic learning process in a neural network. These models exhibit properties that help explain gradual, sometimes punctuated development; graded sensitivity to what a symbolic approach would treat as categorical distinctions; changes in readiness to learn with experience; sensitivity to frequency and typicality, and U-shaped patterns in development.

Findings in the Mathematical Cognition Literature Considered and Modeled from a PDP Perspective

The previous section reviewed models applied to topics outside of mathematical cognition. Are findings in the mathematical cognition literature consistent with the PDP perspective, and can this perspective shed light on their interpretation? In this section we will argue that they are indeed consistent with the overall perspective, and we describe ways in which the perspective is already contributing to an understanding of aspects of mathematical cognition.

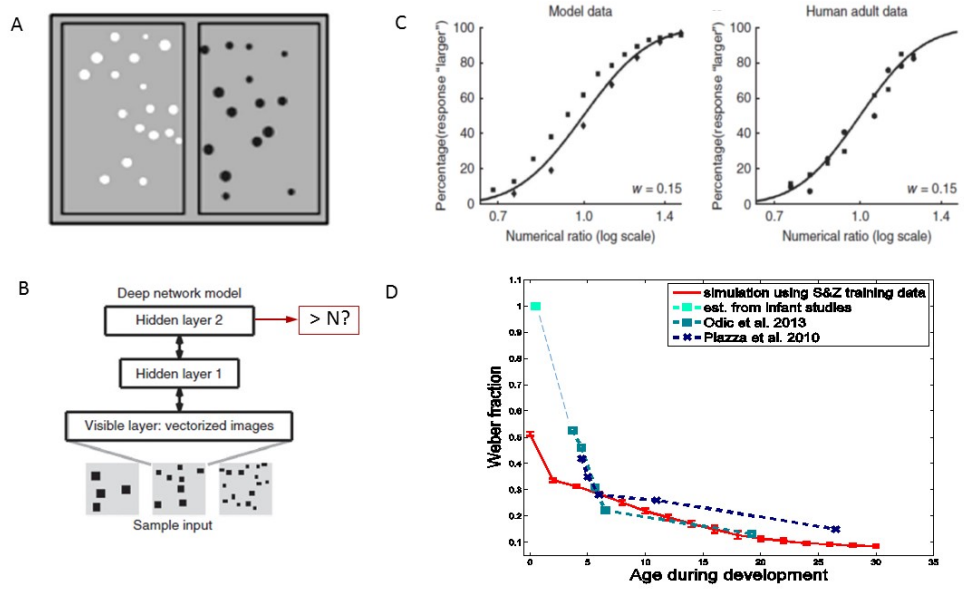


Figure 3. A neural network simulation that captures key characteristics of the approximate number system. (A) Display like that used in many developmental studies of the approximate number system [115–117]. (B) The neural network model used to simulate adult performance in the ANS task [76]. The network is trained to form a representation of its inputs that can be inverted to reproduce the input [118]. Examples of training stimuli illustrating variation in size and number are shown. After network training, a simple classifier is trained to judge whether the number of items in the display is greater than or less than a criterial number (N). (C) Simulation results [76] capturing ratio-dependent sensitivity together with relevant behavioral data [116]. For both model and experiment, similar estimate of the ratio sensitivity index *w* is obtained for data obtained with two different values of N (diamonds: N = 8; squares: N = 16; circles: N = 32). Ratio dependence is indicated by the fact that the classification response (y axis) depends on the ratio of the numerosity of the stimulus presented (x axis) to the numerosity of the standard. (D) Simulation of initial ability and time course of increased ANS sensitivity [77], along with behavioral data replotted from [116,117] with estimate derived from infant studies [119] as estimated by [115,117].

Approximate number. First we consider the concept of an innate Approximate number system (ANS), thought to be shared by species as diverse as rats, pigeons and humans, including human neonates. The ANS exhibits what has been called a signature property [18]: The ability to discriminate differences in numerosity of items scattered in visually-presented arrays depends on the ratio, rather than the difference, of the numerosities. Thus we discriminate 10 dots from 7 as well as we discriminate 100 from 70. While it has been known for quite some time that ANS acuity sharpens with age (neonates discriminate 2:1 ratios, while adults discriminate ratios as small as 8:7) it was previously thought that education had little effect [74]. Recent findings, however, show that uneducated indigenous Amazonian

adults show discrimination ratios more characteristic, on average, of 3-5 year old children, leading to the conclusion that ANS acuity may be education-dependent [75]. Furthermore, recent studies with deep neural network models (Figure 3) show that the representations formed in such networks exhibit ratio-dependent numerosity sensitivity without being trained to represent numerosity per se [76,77]. In fact such models can support initial Weber-law-like sensitivity to numerosity prior to any training, with acuity similar to that of 3-5 year old children [77] and from there they exhibit a gradual increase in acuity as they are trained simply to form a representation that can be used to reconstruct the network's input. The networks and learning procedures are completely generic networks with two layers of modifiable connection weights. Thus the results suggest that evolutionary preparation to represent numerosity *per se,* while it may contribute, may not be necessary for exhibiting signatures of the ANS, and provide a mechanistic basis for the finding that experience affects ANS sensitivity.

Exact number. Proposals that an understanding of exact number is innate were common [78] until the publication of another study in uneducated Amazonian adults [24] whose language has no words for exact number. This study found that these adults failed to create a set of objects with the same cardinality as a given set. While it can be interpreted in other ways [79], the finding is consistent with the view that the lack of number words reflects an underlying lack of a concept of exact number, and other evidence suggests that young children and uneducated indigenous people fail tasks probing basic properties of the exact number concept [25]. Thus, many recent proposals (e.g. [18,80]) exclude a concept of exact number from 'core knowledge' and recent interest has focused on the development of exact numerical ability. As suggested by the quotation from Izard *et al* above, some have sought to characterize development as the discovery or 'semantic induction' of a cardinality principal (CP), indexed by the ability to succeed at the so-called *give-N* task for numbers greater than 4. Indeed, several studies do suggest that children who can succeed for N>4 possess a general procedure (independent of the specific value of N) since, to a first approximation, if they succeed with 5 they also succeed with larger numbers within their known count list. Sarneca and Carey [81] suggested that success at the give-N task for N>4 corresponds to the induction of the Cardinality Principle, and supported this claim by showing that the average performance of a group of children who met the Give-N (N>4) criterion exceeded chance on other tasks thought to reflect possession of the fundamental principle of exact number, the *successor principle*. Concerns with this conclusion were raised in a subsequent paper, however [82]. This study divided children who passed the give-N task for N = 6, 7 and 8 into three groups based on the length of their count list, and showed that children with small count lists (10-19) did not exceed chance in the other tasks even with numbers in the range of 5-7. Children with longer count lists succeeded more often, but their success was more likely for smaller numbers, and did not extend to the largest numbers in their count lists. The authors concluded that a complete understanding of the successor principle arose only after considerable additional experience. From a PDP perspective, we might go a step further and ask whether the characterization of children as possessing a principle as such is the right way to think about their knowledge. Instead we might see this and other principles as approximate characterizations of a system of implicit knowledge that is gradually acquired, and we might view the sense we as adults have of an intuitive understanding of these principles as reflecting an emergent consequence of experience with number. This experience is likely to include discourse about number informed by the efforts of mathematicians to rationalize and formalize number systems [14].

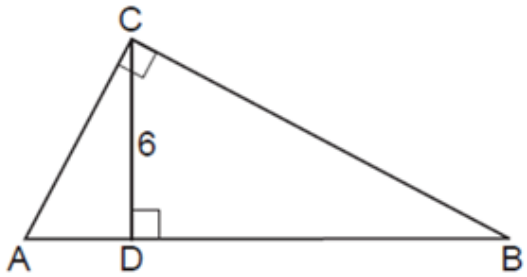
So far we have focused on primitive or at most very basic aspects of mathematical cognition and on the performance of those who are very young and/or have minimal education. However, findings from a wide range of aspects of mathematical cognition are consistent with the view that mathematical skills and abilities emerge slowly and are subject to strong frequency and typicality biases characteristic of the behavior of neural network models that learn gradually from experience (see Box 1[2] for further evidence of such effects and a description of a neural network model that exhibits one of these effects). We highlight here one study that examined understanding and reasoning in geometry across a span from grade school to graduate school [83]. The study noted that even high-school students fail to appreciate the formal structure of mathematical reasoning systems. Noteworthy too is the finding [84] that many high-school and college students who have been repeatedly exposed to basic facts of trigonometry nevertheless make errors that can be seen as reflecting a behavioral tendency driven largely by the statistics of their experiences with mathematical expressions. When given a choice between alternative answers to the question 'which expression has the same value as **cos**(–30°)?' they predominantly choose –**cos**(30°) rather than the correct answer which is **cos**(30°). This response can be seen as an acquired response tendency supported by frequency and typicality, since 'pulling out' a minus sign from inside an expression in parentheses is often correct in an algebraic context. Those who do not make this error have learned instead to treat trigonometric expressions as referring to measurable properties of quantities represented in terms of a conceptual structure that integrates two other conceptual structures, the Cartesian co-ordinate system and a system for representing angles as points on the unit circle [84]. Many students fail to acquire this system, and only those showing evidence of partial prior understanding benefit from a brief lesson explaining it [84] (see also [16]).

In sum, the characteristics of PDP models seem consistent with findings across many domains of mathematics. On this basis, we conclude that it may be worthwhile to see if we can develop the approach further, and also to explore its implications for mathematics teaching and learning.

A Challenge for a 10-year Research Program

Extending the PDP framework to address all aspects of mathematical cognition will certainly be a challenge. For us the extent of the challenge is brought out by a consideration of what it might take for a neural network to do well on a test High-School students must take to demonstrate basic proficiency in Geometry. The quest to build an neural-network based simulation system that could pass one such test (the New York State Regent's exam in Geometry – see Figure 4 for an example problem from the test) was inspired by a challenge issued by the Allen Foundation in 2014, and researchers at the Allen Institute for Artificial Intelligence have recently achieved a partial success [85]. However, the Allen system does not learn – instead it uses knowledge built into it by its programmers and even its developers concede it does not 'understand' geometry. We propose a harder challenge: to teach an artificial agent grounded in our conception of the culturally constructed context in which mathematical abilities are acquired as described above, relying on the conception of cognition and learning embodied

[2] Boxes are located after the bibliography at the end of this document.



In right triangle ABC, CD is the altitude to hypotenuse AB. If CD = 6 and the ratio of AD to AB is 1:5, determine and state the length of BD.

[Only algebraic solutions receive full credit]

Figure 4. An example question from the New York State Regents' exam in Geometry (2014). Students are given a blank page to show their work.

in the PDP framework, and drawing on the technology of artificial neural networks. The goal is that this system should exhibit human like-learning and developmental change, learning from human-like naturalistic and educational experiences, gradually building to the point where it can succeed in passing the test. Because it is a hard challenge, we expect it will require 10 years to succeed in meeting it. We have begun to pursue this path ourselves, but we believe success will depend on the engagement of a larger community; thus we encourage others to join us. Below we sketch a path forward for this effort.

We see three main reasons to pursue this challenge. First, it is clear that neural networks, popular in the 50's and again in the 80's and 90's, are now enjoying a new wave of success as computational models of intelligence, but many still question whether neural networks can really think and reason as humans do. Until very recently most of the clearest successes of neural networks have involved mapping inputs to outputs in a single forward propagation of activation or a single process of settling to a fixed point [55,86] However, many mathematical problems including the one in Figure 4 require a sequence of steps that produce intermediate results needed to solve the problem. We suggest that if a neural network model could learn to solve problems as complex as the one shown in the figure – in a human-like way, when trained with human-like experiences – this would constitute a demonstration that they are capable of really thinking in the same way that humans think. Second, the characteristics of neural networks appear well suited to capturing many aspects of mathematical cognition, as reviewed above, leading to the conclusion that they might provide concrete instantiations of mechanisms that capture the nature of human knowledge and learning within the domain of mathematics, thereby helping us understand in more detail why mathematics is hard to learn, leading ultimately to improved practices for learners and teachers.

Third, addressing the challenge will surely require further development of the neural network models themselves, both conceptually and technically. As we will review briefly below, there has been an explosion of recent developments in neural network architectures. Pursuing the challenge we propose should help direct these developments toward capturing as-yet uncaptured aspects of the nature of human intelligence.

Table 1 – Tenets of the PDP approach to Mathematical Cognition

- Mathematics offers **culturally constructed model systems** that support reasoning in number, geometry, and many other domains
- **Some characteristics** of such systems are **implicit in everyday experience** interacting with objects in the physical world
- Others arise from **structured interactions with physical instantiations of these models**, or with symbols that are grounded in such models, **guided by peers, caregivers, and teachers**
- **Neural networks have many of the right characteristics** to capture how humans acquire an understanding of such models, **but need to be extended** to succeed.

The basic tenets of the proposed approach (Table 1) arise from the arguments made above. Below, we provide an overview of our planned approach, and then review recent developments from research on artificial neural networks for machine learning that we believe will be helpful as we proceed.

The project will focus on an artificial agent that lives in a simulated two-dimensional world (Figure 5) that it will observe through an eye it can move to fixate at any location [87]. Its input from this world will be a bitmap centered at the eye's point of fixation. The agent will be allowed to explore and manipulate its world using a simple simulated hand that it can use to touch, drag, flip, and rotate movable objects it encounters in its environment, and it will have a separate character-based input-output system through which it can receive and produce language (typed sequences of characters) and mathematical expressions. The agent will learn through observation of naturalistic events that may take place in its environment, including the movements of simple animats that may move through the environment, and through its own exploration. T will also learn through structured experiences guided by supportive teaching inputs, simulating the roles of caregivers, peers, and classroom teachers.

Initial exploration should allow the agent to acquire, among other things, intuitions of invariance and conservation of shape and number like those underlying the proof of the Pythagorean Theorem in Figure 1, on which structured experience will then build. Over the course of development, we will expect the agent to learn to act in response to linguistic inputs, to use instructions, demonstrations, and explanations presented in language, and even to explain and justify its actions. At later stages of development, the agent's simulated world will contain tools that will allow it to measure, copy and construct geometrical diagrams and graphs according to procedures specified by teaching inputs (much as Euclidean geometry provides instructions in how to construct figures with specified characteristics using compass and straightedge). Through such experiences the agent will progress through a curriculum in mathematics similar to that in a good school system in an advanced country, allowing it to acquire intuitions and skills related to number, arithmetic, measurement, algebra, and geometry. This idea is similar to the curriculum learning approach frequently explored in contemporary machine learning [88]. Through this we propose that the agent should be able to acquire human-like abilities to reason mathematically and to carry out multi-step mathematical problem solving activities, under the

continued guidance of its teachers, up to and including all of the plane geometry questions on the New York State Regents exam, including problems like the one in Figure 4, as well as other problems that require justification and explanation.

The framework outlined above follows in the tradition of past PDP models of seeking to keep the model as simple as possible while still allowing it to exemplify essential principles and to explore essential issues [2,89]. The model is embodied in its simple 2-dimensional world, which is sufficient for it to have experience with, for example, the relationship between time and space and of the characteristics of paths of objects as they move through space [90]. Yet it avoids a wide range of issues that may not be essential to the goals of the project that would arise if the agent were actually embodied in a physical robot. Of course it is possible that there are important aspects of embodied experience that the agent will lack – what these may be should become apparent as the project develops. Similarly, the language input and output of the model will be in the form of character strings, allowing us to avoid having to implement systems for spoken and hand-written language processing so as to allow focus on exploring the agent's ability to relate linguistic and symbolic mathematical expressions to properties and relations among quantifiable properties of objects and sets of objects presented in its 2-dimensional world. The approach is similar to the one taken in [57,87] in teaching a simulated agent to learn to play Atari games and to deploy attention to objects at different locations in space.

A part of the work of the project will be to specify the architecture of the neural network, allowing exploration of just want really needs to be built in. Depth is now understood to be important, and some modularization at lower levels of input-output processing for linguistic vs. visuospatial inputs seems warranted by basic facts about the organization of the brain. Beyond this, our own approach would be to build in as little as possible. However, in keeping with the complementary learning systems theory of the organization of learning systems in the brain, we do expect the agent to require a hippocampus-like fast learning system to complement the slow-learning deep neural networks that will constitute its analog of neocortex [91,92].

Key Aspects of the Challenge and how we Plan to Address Them

Prototypical neural network models may capture some aspects of human abilities, but they are still quite limited in some ways. Consider, for example, the highly successful deep convolutional neural network models of visual object recognition. They perform very well at classifying objects found in still images (though still not quite at human levels) and capture patterns of neural activity observed in primate and human participants when processing the same displays that the models process [55,93,94]. They do so simply by learning a set of connection weights from a training set that projects from relatively unprocessed inputs onto a set of classification units. These models lack several essential characteristics for success at mathematical cognition. Importantly, though, a considerable body of work dating from the late 1980's has also recently begun to see success in overcoming many of the more important limitations of these particular types of neural networks. We consider four of the limitations of deep CNN models that seem particularly relevant for mathematical cognition and discuss recent developments that encourage the view that these limitations are now being overcome, supporting the idea that it may be possible to make progress addressing the stated challenge.

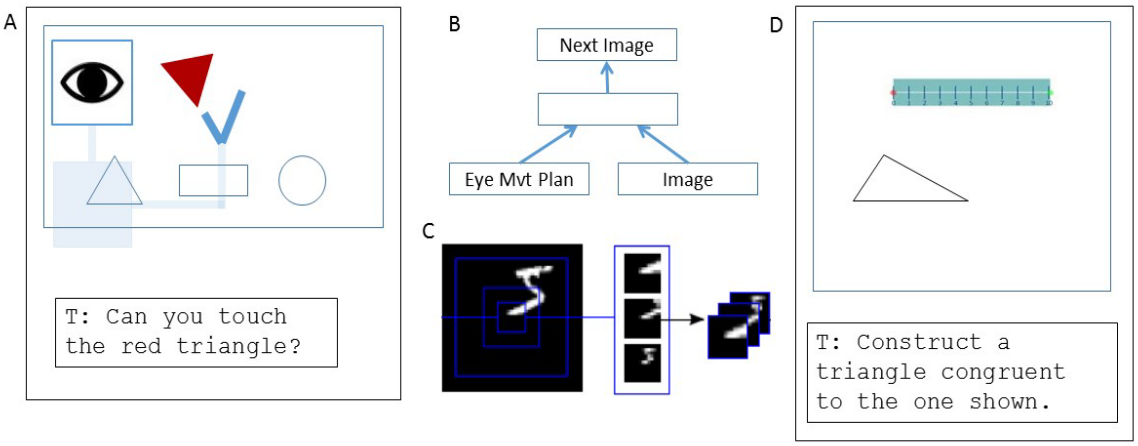


Figure 5. (A) A visual display that might be presented to the agent, showing a possible position of the agent (eye and V-shaped 2-fingered hand highlighted) positioned over the display. Also shown is a possible verbal instruction to the Agent. Verbal input could also request a verbal response, such as 'What is the shape of the red object?' The agent could use the transforming auto-encoder network (B) [96] to learn to anticipate what an object in the display would look like once fixated. The agent's visual input would be a multi-resolution image like that used in [87] (C), and would change as the agent moved its point of fixation. Later in its development, the agent could be asked to carry out constructions, using tools such as the ruler that it will learn to manipulate with its hand. The tool could be a Euclidean straightedge without markings, or it could allow for measurement by having markings at equal intervals, as in the example shown.

Combining supervised learning with unsupervised and exploration-based learning. While CNNs have been trained primarily using a supervised learning approach, it seems likely that some of the basic intuitions underlying mathematical cognition arise from experience interacting with objects in the world, independent of the guidance of teachers. Some of these experiences may be very general, but still lead to the development of early schemas that are then applicable to mathematical cognition [90]. *Unsupervised* neural network training regimes may contribute importantly to the development of these intuitions [76]. In unsupervised learning, a network is trained simply to form an internal representation of presented inputs that is sufficient to reproduce the input itself. Hinton pioneered unsupervised learning in the 1980's and has since developed the approach extensively [95], and as noted above, neural network models trained with unsupervised learning acquire representations capturing key characteristics of the approximate number system. An extension of these ideas may help our agent learn invariant representations of shape. Consider the situation in which the agent is fixated at one location in its world, with an object present in another location (Figure 4a). Hinton [96] has proposed a *transforming auto-encoder* system for learning an invariant representation of the objects shape in this situation (Figure 5b). Suppose the agent can construct (before acting on it) a pattern of activation corresponding to a plan to move its point of fixation to the object. This pattern, together with its

current retinal image, would allow the agent to anticipate what the object will look like after the eye movement. Early in learning, this anticipated pattern might be based on random initial connection weights, but the mismatch between the anticipated pattern and the one actually observed after the eye movement could drive learning. In this way, the network would gradually learn to represent in anticipation what objects in the periphery will look like once brought to the point of fixation, allowing such representations to serve as canonical (and location-invariant) object representations. The idea could be extended to other transformations of the object, such as those produced if the agent dragged, rotated, or flipped it with its hand.

Discovering strategies based on deep reinforcement learning. Classification [55] and translation systems [56] depend ultimately on training examples consisting of a corpus of input and output pairs, and some of mathematics learning might be characterized as learning from such examples. However, human learners innovate, finding new procedures never modeled by their teachers [30]. The methods of reinforcement learning [97] provide a means for capturing these abilities, since reinforcement learning progresses in part through exploration of the space of possible actions. For example, the TDgammon program used reinforcement learning to find novel approaches to win at Backgammon [54], and the neural network that has learned to play Atari games [57] uses reinforcement learning together with a multi-layer neural network to discover clever strategies that allow it to perform at or above human performance levels on many games. Reinforcement learning can also proceed using internally-generated reward signals, for example based on the novelty of an experience that can facilitate later learning when an external source of reinforcement (created by a task set by an external teacher) arises [98]. Indeed the unsupervised learning described in the preceding paragraph might be driven by novelty-seeking, and would be implemented as such in our learning agent. It is likely that it will continue to be productive to integrate exploration-based unsupervised learning, reinforcement learning and supervised learning, similar to the approach that has produced expert level performance in the demanding game of Go [99].

Processing structured inputs and producing structured sequential behavior. The objects of mathematical thought are often highly structured, whether they be diagrams like the ones in Figures 1 and 4 or symbolic expressions, and critiques of neural network-based approaches have often focused on their apparent insensitivity to such structure. However, neural networks sensitive to sequential structure in linguistic inputs have been around since the late 1980s [100]. These networks appeared to be limited in their ability to learn over sufficiently long stretches of context to deal effectively with long-distance dependencies, and so were not in favor in machine learning for many years. However, an effective solution to the long-distance dependency problem was introduced in 1997 [101] and recently this solution has been exploited in machine translation systems that are now the state of the art [56]. Such systems process input sentences with natural language structure and produce appropriately structured outputs in the target language, without an explicitly structured internal representation. Also quite recently [102], a system based on these ideas have been used to create a neural Turing machine – a system that exploits system-internal short term memory banks organized in a linear array, similar to the storage cells in the tape of the original Turing Machine.

The NTM is an exciting development that should be built on training in our agent, with a key difference: Instead of a system-internal memory, we propose that the agent should work with linearly-organized items placed in its two-dimensional workspace. This would include arrays of objects or numbers presented in its two-dimensional world. For example the agent could learn to count a row of objects or add a column of numbers. In this setting, the agent's point of fixation in its world would correspond to the position of the Turing Machine's read-write head in its tape, or the pointer in the NTM to the current item it is processing in its internal memory array. This will allow the agent's teachers to present it with problems to solve, demonstrate problem solutions to the agent, and observe the agent's actions as it engages in problem solving behavior [29], all using the world rather than the NTM's internal memory, an approach to teaching and learning mathematics now actively being explored [28]. We have begun to use this approach in initial investigations of teaching a simple instantiation of our agent to count (see Box 2). The teacher can place several tokens in the agent's world and demonstrate how to count them; it can then encourage the agent to do the same and observe its actions as it proceeds, allowing monitoring and feedback not only on the outcome of the agent's counting process but on each of the steps along the way. This setting also provides a framework for teaching the agent to engage in a structured sequence of activities (for example, a small number addition procedure taught in some pre-schools assembles previously-acquired counting-related procedures, see Box 2).

Integrating language and visuospatial cognition. Our goals for the agent include the ability to solve problems presented in verbal form (or with verbal instructions accompanying a diagram), to follow instructions and understand explanations, and even to produce explanations and justifications for its problem solving steps. In our view, knowledge in connections underlies the ability to understand or produce overt propositional language, and to carry out structured activities such as mathematical problem solving in response to instructions. Propositional statements that a person might utter are thus not thought of as encoded directly as such, but as arising from a dynamic, activation-based process that results in the construction of these utterances guided by knowledge in connections. Under this conceptualization, there is not a simple or transparent relationship between our cognitive abilities and our verbalizable or otherwise consciously accessible knowledge [103,104], and we hold the view that learning plays an important role in establishing those links that do exist [70]. That said, we do acknowledge that there must be some interaction between propositionally based and visuospatial and enactive aspects of mathematical abilities. Achieving the goal of capturing this interaction will require the integration of language and visuospatial cognition – and many still see this as an insurmountable hurdle for an approach that avoids building in explicit symbolic processing capabilities [105]. Work in the PDP framework, however, has argued that avoiding building in such capabilities is an advantage for language processing and other systematic cognitive tasks [106,107], and several recent deep-neural network approaches in machine learning are achieving state of the art results in language interpretation tasks and mapping from images to language using learned distributed representations rather than structured symbolic representations [56,108]. The challenge we face to allow a sufficient degree of interactive engagement between linguistic and visuospatial reasoning processes such that verbal statements can guide, explain, and justify action will not be easy to address, but we see it as essential to the ultimate success of our endeavor, as we view capturing human-level abilities to benefit from verbal

input as well as to explain and justify their own actions as important parts of a complete model of mathematical cognition and of expert human cognitive abilities in any other cognitive domain.

In summary, the discussion above points to some of the challenges we face in our effort to create a simulated agent that could acquire mathematical ability in a human like way, sufficient to take the New York state Regents' exam in Geometry. While we have pointed to recent progress that provides a starting place for this effort, we do not intend to suggest that meeting the challenge will be easy. Building on new developments in neural networks for machine learning and artificial intelligence, we are hopeful that over the next 10 years it will be possible to make real progress toward this goal (See Box 3, Outstanding Questions).

Implications for teaching and learning

The PDP approach to mathematical cognition has many implications for learners, teachers, and the activities of learning and teaching – implications that relate to the procedures and narratives that are employed in the teaching and learning of mathematics. We focus on three essential points.

From the point of view of the learner, an essential observation is that *understanding emerges gradually from experience.* We believe it is important for all learners to understand that the initial feeling of incomprehension they may experience when encountering a new domain of mathematics is natural any typical, should not be seen as a sign of an inherent lack of ability. Learners need to understand that comprehension will arise as they gain experience and facility. A corollary of this is a rebalancing or our conception of the basis for outstanding accomplishment in mathematics: Perhaps the genius of great mathematicians comes as much from their propensity to engage in mathematics-relevant experiences and activities as much from any inherent abilities. Of course this propensity itself is affected by social and cultural context, and excitement and encouragement as well as the opportunity to build on the knowledge of others are all likely to be essential factors. Ensuring all developing children benefit from frequent exposure to experiences encouraging the development of mathematical intuitions will be a key factor in fostering mathematical ability.

From the point of view of the teacher, an essential observation is that what is hard for the student may not be so easy for the teacher to discern. The habits of mind we acquire from experience, once they become automatic, result in an intuition of self-evidence that the untutored mind does not share. Signs of this are found in many places in mathematics education. The Platonic precept that mathematical ideas are universal encourages the illusion that the explanation the teacher understands is transparent to the student, and the illusion that operations and procedures that seem obvious to the teacher will be easy for the student to understand. Indeed research shows that teachers are often wrong in their beliefs about what will be hard or easy for students [109]. Effective teaching may depend on finding ways of letting students go through the gradual process of acquiring experiences that constitute the substrate of intuitive understanding, rather than expecting them to follow the thought processes and reasoning patterns of an expert.

Finally, we consider the experiences and activities of learning and teaching. An approach to mathematics that emphasizes the role of culture and experience requires and understanding of the

learning environment and the interaction of learners with their environments. We need to know, not only for our modeling project, but also for an understanding of how real children and scholars learn, what the cultural practices are that structure the interactions between learners and teachers; how these vary within and across cultures; and how the teacher responds to the learner's progress and explorations. Others have stressed the importance of understanding how what children learn depends on their experiences in learning contexts [110], but the issue is often reduced to simple environmental occurrence frequencies over items or problem types [111,112]. While understanding experience frequency is certainly an important first step, to go forward we will need a more thorough characterization of all aspects of the situations in which learners gain experiences by interacting with their physical, social, cultural, and educational environment.

The perspectives we have articulated in this section are not unique to us – scholars of teaching and learning mathematics have long made similar points [83,109]. The important point is that the PDP framework provides a theoretical framework within which these observations can be understood as inherent in the nature of the human mind, and a growing set of tools for explicitly capturing them in computational models that learn. As the framework develops over the next 10 years, we hope it will contribute to the development of a firm scientific framework in which these ideas can be further explored.

Bibliography

1 Rumelhart, D.E. *et al.* (1986) *Parallel distributed processing: Explorations in the microstructure of cognition (V1 and V2)*, MIT press.

2 Rogers, T.T. and McClelland, J.L. (2014) Parallel distributed processing at 25: Further explorations in the microstructure of cognition. *Cogn. Sci.* 38, 1024–1077

3 Russell, B. (1903) *The principle of mathematics*, Cambridge University Press.

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

5 Martin, W.A. and Fateman, R.J. (1971) , The MACSYMA system. , in *Proceedings of the Second Symposium on symbolic and algebraic manipulation.*, pp. 59–75

6 Needham, T. (1997) *Visual complex analysis*, Clarendon Press.

7 Shepard, R.N. (2008) The Step to Rationality : The Efficacy of Thought Experiments in Science , Ethics , and Free Will. 32, 3–35

8 Greeno, J.G. (1991) Number sense as situated knowing in a conceptual domain. *J. Res. Math. Educ.* 22, 170–218

9 Shepard, R.N. and Metzler, J. (1971) Mental Rotation of Three-Dimensional Objects. *Science (80-.).* 171, 701–703

10 Wallace, E.C. and West, S.F. (2015) *Roads to geometry*, Waveland Press.

11 Thompson, P.W. (1993) Quantitative reasoning, complexity, and additive structures. *Educ. Stud. Math.* 25, 165–208

12 Barwise, J. and Etchemendy, J. (1991) , Visual information and valid reasoning. , in *Visualization in teaching and learning mathematics*, pp. 9–24

13 Hirsh, R. (1997) *What is mathematics, really?*, Oxford University Press.

14 MacLane, S. (1986) *Mathematics: Form and function*, Springer-Verlag.

15 Lamb, W.R.M. (1924) *Plato: Laches; Protagoras; Meno; Euthydemus*, 2Loeb Classical Library.

16 Goldin, A.P. *et al.* (2011) From ancient Greece to modern education: Universality and lack of generalization of the socratic dialogue. *Mind, Brain, Educ.* 5, 180–185

17 Kant, I. and Guyer, P. (1998) *Critique of pure reason*, Cambridge University Press.

18 Feigenson, L. *et al.* (2004) Core systems of number. *Trends Cogn. Sci.* 8, 307–314

19 Brannon, E.M. and Merritt, D.J. (2011) *Evolutionary Foundations of the Approximate Number System*, 1Elsevier Inc.

20 Leslie, A.M. *et al.* (2008) The generative basis of natural number concepts. *Trends Cogn. Sci.* 12, 213–8

21 Butterworth, B. (2010) Foundational numerical capacities and the origins of dyscalculia. *Trends Cogn. Sci.* 14, 534–41

22 Cole, M. and Scribner, S. (1974) *Culture and thought: A Psychological Introduction*, John Wiley & Sons.

23 Vygotsky, L.S. (1978) *Mind in Society: The Development of Higher Psychological Processes*, 1

24 Gordon, P. (2004) Numerical Cognition without Words: Evidence from Amazonia. *Science (80-.).* 306, 496–499

25 Izard, V. *et al.* (2008) Exact Equality and Successor Function: Two Key Concepts on the Path towards understanding Exact Numbers. *Philos. Psychol.* 21, 491

26 Izard, V. *et al.* (2014) Toward exact number: Young children use one-to-one correspondence to measure set identity but not numerical equality. *Cogn. Psychol.* 72, 27–53

27 Menninger, K. (1969) *Number words and number symbols: A cultural history of numbers.*, MIT Press.

28 Alibali, M.W. and Nathan, M.J. (2012) Embodiment in Mathematics Teaching and Learning: Evidence From Learners' and Teachers' Gestures. *J. Learn. Sci.* 21, 247–286

29 Alibali, M.W. and DiRusso, A. a. (1999) The function of gesture in learning to count: more than keeping track. *Cogn. Dev.* 14, 37–56

30 Siegler, R.S. and Jenkins, E. (1989) How children discover new strategiesErlbaum. *Hillsdale, NJ*

31 Case, R. and Okamoto, Y. (1996) No Title. *Mongraphs Soc. Res. Child Dev.* 61,

32 Dehaene, S. (1992) Varieties of Numerical Abilities. *Cognition* 44, 1–42

33 Siegler, R.S. and Lortie-Forgues, H. (2014) An Integrative Theory of Numerical Development. *Child Dev. Perspect.*

34 Dehaene, S. and Cohen, L. Towards an anatomical and functional model of number processing. , *Mathematical Cognition*, 1. (1995) , 83–120

35 Dehaene, S. *et al.* (1993) The mental representation of parity and number magnitude. *J. Exp. Psychol. Gen.* 122, 371–396

36 Nunez, R. (2011) No Innate Number Line in the Human Brain. *J. Cross. Cult. Psychol.* 42, 651–668

37 Opfer, J.E. and Siegler, R.S. (2007) Representational change and children's numerical estimation. *Cogn. Psychol.* 55, 169–195

38 Barth, H.C. and Paladino, A.M. (2011) The development of numerical estimation: evidence against a representational shift. *Dev. Sci.* 14, 125–135

39 Link, T. *et al.* (2014) Unbounding the mental number line—new evidence on children's spatial representation of numbers. *Front. Psychol.* 4, 1–12

40 Ramani, G.B. *et al.* (2012) Taking It to the Classroom : Number Board Games as a Small Group

Learning Activity. 104, 661–672

41 Margolis, H. (1993) *Paradigms and Barriers*, The University of Chicago Press.

42 Fodor, J.A. (1975) *The language of thought*, 4

43 Fodor, J.A. (2010) *LOT 2: The Language of Thought Revisited*, 9780199548

44 Spelke, E.S. *et al.* (1992) Origins of Knowledge. *Psychol. Rev.* 99, 605–632

45 Carey, S. (1991) Knowledge Acquisition: Enrichment or Conceptual Channge? In *The Epigenesis of mind* (Carey, S. and Gelman, R., eds), pp. 257–291, Erlbaum

46 Pinker, S. (1999) *Words and rules: The ingredients of language*, Basic Books.

47 Rumelhart, D.E. *et al.* (1986) Learning representations by back-propagating errors. *Nature* 323, 533–536

48 Elman, J.L. *et al.* (1996) *Rethinking Innateness: Connectionist Perspective on Development*,

49 Linsker, R. (1986) (2)From basic network principles to neural architecture: emergence of spatial-opponent cells. *Proc. Natl. Acad. Sci. U. S. A.* 83, 7508–7512

50 Rumelhart, D.E. and McClelland, J.L. (1986) On learning the past tenses of English verbs. In *Parallel distributed processing: explorations in the microstructure of cognition, vol. 2* (McClelland, J. L. and Rumelhart, D. E., eds), pp. 216–271

51 Pinker, S. and Prince, A. (1988) On launguage and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition* 28, 73–193

52 Plaut, D.C. *et al.* (1996) Understanding normal and impaired word reading: computational principles in quasi-regular domains. *Psychol. Rev.* 103, 56–115

53 Seidenberg, M.S. and Plaut, C. (2014) Quasiregularity and Its Discontents : The Legacy of the Past Tense Debate. 38, 1190–1228

54 Tesauro, G. (1994) TD-Gammon, a self-teaching backgammon program, achieves master-level play. *Neural Comput.* 6, 215–219

55 Krizhevsky, A. *et al.* (2012) Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.*

56 Sutskever, I. *et al.* (2014) , Sequence to Sequence Learning with Neural Networks. , in *Neural Information Processing Systems 2014 Proceedings*, pp. 1–9

57 Mnih, V. *et al.* (2015) Human-level control through deep reinforcement learning. *Nature* 518, 529–533

58 McClelland, J.L. and Jenkins, E. (1991) Nature, Nurture, & Connections: Implications of connectionst models for cognitive development. In *Arhcitectures for Intelligence* (van Lehn, K., ed), pp. 41–73, Erlbaum

59 Siegler, R.S. (1976) Three aspects of cognitive development. *Cogn. Psychol.* 8, 481–520

60 McClelland, J.L. (1995) A connectionist account of knoweldge and development. In *Developing cognitive competence: Noew approaches to process modeling* (Simon, T. J. and Halford, G. S., eds), pp. 157–204, Lawrence Erlbaum Associates

61 Schapiro, A.C. and McClelland, J.L. (2009) A connectionist model of a continuous developmental transition in the balance scale task. *Cognition* 110, 395–411

62 Jansen, B.R.J. and Van der Maas, H.L.J. (2001) Evidence for the Phase Transition from Rule I to Rule II on the Balance Scale Task. *Dev. Rev.* 21, 450–494

63 Saxe, A.M. *et al.* (2013) , Learning hierarchical category structure in deep neural networks. , in *Proceedings of the 35th annual meeting of the cognitive science sociiety*, pp. 1271–1276

64 Henik, A. and Tzelgov, J. (1982) Is three greater than five: The relation between physical and semantic size in comparison tasks. *Mem. Cognit.* 10, 389–395

65 Girelli, L. *et al.* (2000) The development of automaticity in accessing number magnitude. *J. Exp. Child Psychol.* 76, 104–22

66 Dehaene, S. *et al.* (1993) The mental representation of parity and number magnitude. *J. Exp. Psychol. Gen.* 122, 371

67 Fischer, M.H. (2008) Finger counting habits modulate spatial-numerical associations. *Cortex* 44, 386–392

68 Cohen, J.D. *et al.* (1990) On the control of automatic processes: A parallel distributed processing account of the Stroop effect. *Psychol. Rev.* 97, 332–361

69 MacLeod, C.M. and Dunbar, K. (1988) Training and Stroop-like interference: evidence for a continuum of automaticity. *J. Exp. Psychol. Learn. Mem. Cogn.* 14, 126

70 Rogers, T.T. and Mcclelland, J.L. (2003) Semantic Cognition : A Parallel Distributed Processing Approach.

71 Gelman, R. and Williams, E.M. (1998) Enabling constraints for cognitive development and learning: Domain specificity and epigenesis. In *Handbook of Child Psychology* 1pp. 575–630

72 Chapman, K.L. and Mervis, C.B. (1989) Patterns of object-name extension in production. *J. Child Lang.* 16, 561–571 ST – Patterns of object–name extension in

73 Rogers, T.T. *et al.* (2004) U-shaped curves in development: A PDP approach. *J. Cogn. Dev.* 5, 137–145

74 Pica, P. *et al.* (2004) Exact and approximate arithmetic in an Amazonian indigene group. *Science* 306, 499–503

75 Piazza, M. *et al.* (2013) Education enhances the acuity of the nonverbal approximate number system. *Psychol. Sci.* 24, 1037–1043

76 Stoianov, I. and Zorzi, M. (2012) Emergence of a “visual number sense” in hierarchical generative models. *Nat. Neurosci.* 15, 194–196

77 Zou, Y. *et al.* (2015) Initial competence and developmental refinement of a sense of number in a deep neural network. *submitted.*

78 Carey, S. and Gelman, R. (1991) *The Epigenesis of Mind: Essays on Biology and Cognition*, LEA.

79 Frank, M.C. *et al.* (2008) Number as a cognitive technology: Evidence from Pirahã language and cognition. *Cognition* 108, 819–824

80 Feigenson, L. and Carey, S. (2005) On the limits of infants' quantification of small object arrays. *Cognition* 97, 295–313

81 Sarnecka, B.W. and Carey, S. (2008) How counting represents number : What children must learn and when they learn it q. 108, 662–674

82 Davidson, K. *et al.* (2012) Does learning to count involve a semantic induction? *Cognition* 123, 162–173

83 Burger, W.F. and Shaughnessy, J.M. (1986) Characterizing the van Hiele Levels of Development in Geometry. *J. Res. Math. Educ.* 17, 31–48

84 Mickey, K.W. and McClelland, J.L. (2016) Understanding Trigonometry as a Coherent Conceptual System.

85 Seo, M. *et al.* (2015) Solving Geometry Problems: Combining Text and Diagram Interpretation. *EMNLP*

86 Salakhutdinov, R. and Hinton, G. (2012) An Efficient Learning Procedure for Deep. 2006, 1967–2006

87 Mnih, V. *et al.* (2014) Recurrent Models of Visual Attention. *arXiv Prepr. arXiv1406.6247* at <http://arxiv.org/abs/1406.6247>

88 Bengio, Y. *et al.* (2009) , Curriculum learning. , in *Proceedings of the 26th annual international conference on machine learning*, pp. 41–48

89 Mcclelland, J.L. (2009) The Place of Modeling in Cognitive Science. *Top. Cogn. Sci.* 1, 11–38

90 Lakoff, G. and Nunez, R. (2000) *Where Mathematics Comes From: How the Embodied Mind Brings Mathematics Into Being*, Basic Books.

91 McClelland, J.L. *et al.* (1995) WHY THERE ARE COMPLEMENTARY LEARNING-SYSTEMS IN THE HIPPOCAMPUS {AND} NEOCORTEX - INSIGHTS FROM THE SUCCESSES {AND} FAILURES OF CONNECTIONIST MODELS OF LEARNING {AND} MEMORY. *Psychol. Rev.* 102, 419–457

92 Kumaran, D. *et al.* What learning systems do intelligent agents need? Complementary Learning Systems Theory Updated. *Trends Cogn. Sci.*

93 Yamins, D.L.K. *et al.* (2014) Performance-optimized hierarchical models predict neural responses in higher visual cortex. DOI: 10.1073/pnas.1403112111

94 Khaligh-Razavi, S.-M. and Kriegeskorte, N. (2014) Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLoS Comput. Biol.* 10, e1003915

95 Hinton, G.E. (2007) Learning multiple layers of representation. *Trends Cogn. Sci.* 11, 428–434

96 Hinton, G.E. *et al.* (2011) Transforming auto-encoders. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* 6791 LNCS, 44–51

97 Sutton, R.S. and Barto, A.G. (1998) Introduction to Reinforcement Learning. *Learning* 4, 1–5

98 Singh, S. and Barto, A.G. Intrinsically Motivated Reinforcement Learning.

99 Silver, D. *et al.* (2016) Mastering the game of Go with deep neural networks and tree search. *Nature* 529, 484–489

100 Elman, J.L. (1990) Finding structure in time. *Cogn. Sci.* 14, 179–211

101 Hochreiter, S. and Schmidhuber, J. (1997) Long Short-Term Memory. *Neural Comput.* 9, 1735–1780

102 Graves, A. *et al.* (2014) Neural Turing Machines. *Arxiv* at <http://arxiv.org/abs/1410.5401>

103 Cleeremans, A. (2014) Connecting Conscious and Unconscious Processing. *Cogn. Sci.* 38, 1286–1315

104 Karmiloff-Smith, a (1986) From meta-processes to conscious access: evidence from children's metalinguistic and repair data. *Cognition* 23, 95–147

105 Tenenbaum, J.B. *et al.* (2011) How to grow a mind: statistics, structure, and abstraction. *Science* 331, 1279–1285

106 St. John, M.F. and McClelland, J.L. (1990) Learning and applying contextual constraints in sentence comprehension. *Artif. Intell.* 46, 217–257

107 McClelland, J.L. *et al.* (2010) Letting structure emerge: Connectionist and dynamical systems approaches to cognition. *Trends Cogn. Sci.* 14, 348–356

108 Vinyals, O. *et al.* (2014) Show and Tell: A Neural Image Caption Generator. *arXiv* DOI: 10.1109/CVPR.2015.7298935

109 Nathan, M.J. (2012) Rethinking Formalisms in Formal Education. *Educ. Psychol.* 47, 125–148

110 Klahr, D. (2012) Patterns, Rules, and Discoveries in Life and Science. In *The Journey From Child to Scientist: Integrating Cognitive Development and the Education Sciences* (Carver, S. and Shrager, J., eds), American Psychological Association

111 Piantadosi, S.T. A rational analysis of the approximate number system. *Psychon. Bull. Rev.*

112 Capraro, R.M. *et al.* (2012) Changes in equality problem types across four decades in four second and sixth grade textbook series. *J. Math. Educ.* 5, 166–189

113 McNeil, N.M. (2007) U-shaped development in math: 7-year-olds outperform 9-year-olds on equivalence problems. *Dev. Psychol.* 43, 687–695

114 Mickey, K.W. and McClelland, J.L. (1998) , A neural network model of learning mathematical equivalence. , pp. 1012–1017

115 Halberda, J. and Feigenson, L. (2008) Developmental change in the acuity of the “number sense”: The approximate number system in 3-, 4-, 5-, and 6-year-olds and adults. *Dev. Psychol.* 44, 1457–1465

116 Piazza, M. *et al.* (2010) Developmental trajectory of number acuity reveals a severe impairment in developmental dyscalculia. *Cognition* 116, 33–41

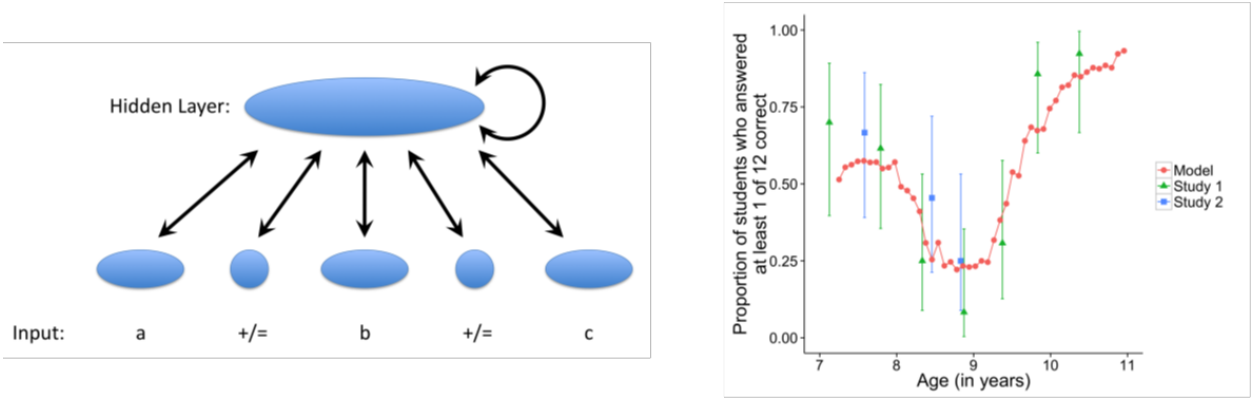
117 Odic, D. *et al.* (2013) Developmental Change in the Acuity of Approximate Number and Area Representations. 49, 1103–1112

118 Hinton, G.E. *et al.* (2006) A fast learning algorithm for deep belief nets. *Neural Comput.* 18, 1527–54

119 Xu, F. and Spelke, E.S. (2000) Large number discrimination in 6-month-old infants. 74, 1–11

Box 1. Capturing U-shaped developmental trends in arithmetic learning

An example of a typicality-dependent U-shaped developmental trend is seen in children's answers to a subset of so-called 'equivalence' problems, in which the child must say what number goes in the blank in an expression such as '5 + 4 + 3 = 5 + __'. In two studies ([113], Figure i, right) less than 10% of 9 year olds correctly answered any of twelve such problems, and a frequent error was to add all the numbers. Further research revealed that the examples children practiced rarely had operands to the right of the equal sign, with '5 + 4 + 3 + 5 = __' being a very frequent problem type. A simple neural network (Figure i, left) with recurrent connections [114] captured this and other features of the data (Figure i, right). Importantly, as training progressed, the network learned its way out of the hole it had dug itself into, even though only a small fraction of the training examples had operands on the right of the equal sign. This learning pattern results from the fact that the learning in these systems is error-driven. At first, error can be reduced quickly by adopting a simple strategy that works for most but not all problems; eliminating the error that remains then involves becoming sensitive to the exact placement of the equal sign and plus signs, over-riding the simple strategy when they are arranged atypically. The result after sufficient learning is a system that behaves in accordance with the principle that the sum of the numbers on both sides of the equal sign should be equal.



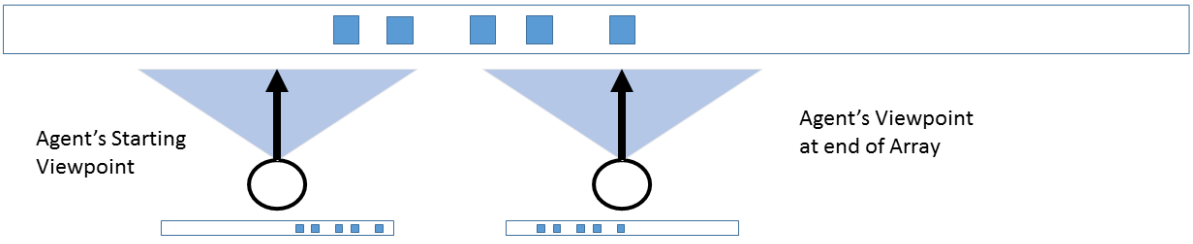
Box 1 Figure i. Left, the recurrent neural network used in [114]t o simulate learning and responding to equivalence problems. Right, the network's performance over the course of its learning on problems of the form *a* = *b* + _ at different points in training. Data from two experiments in [113] are shown for comparison. The network's performance (averaged over simulation runs, in red) has been shifted and scaled to best match the pattern of behavioral data.

Box 2: Teaching a simulated agent to count and to add

To give a flavor for our overall approach, we describe here how a simulated agent might learn to count a linear array of objects (the 'How Many' task). This agent lives in a one-dimensional environment, into which its teacher can place objects for the agent to count. The agent has a head, eye, and hand which for simplicity move in lock-step so that it can simultaneously look at and touch objects. The task of the agent is to touch each object exactly once, saying aloud the next number in the count list. A critical feature of the approach (different from other models) is that the agent's input changes as it proceeds to count. At first all of the objects in the display are to the right of its fixation, and when the count is complete there are no more objects to the right (Figure i). Thus the agent can learn a simple stopping rule. This approach promotes generalization to novel arrangements of inputs (there are over 68,000 different inputs containing 7 items).

The model can be extended to the 'give-N' task by allowing the agent to drag tokens from a stash to positions in the display, under the guidance of the teacher. This time some form of memory is required, making the task harder. Crucially, with mastery of the 'how-many' and 'give-N' tasks, the teacher can guide the agent to learn to add, following a procedure explicitly taught in some preschools [30]. The task of adding 2+3 is broken down into a give-2, a give-3, and a how many. Thus the hierarchical structure of the task is conveyed to the agent by its cultural/educational environment.

The model provides a concrete context in which to explore a wide range of issues, among them the details of the neural network architecture, the particular choice of learning algorithm, and many aspects of the teacher's policy, including the distribution of items to use to train the agent, how much the teacher should demonstrate (which can be simulated for the how-many task by guiding the agent's hand and reciting the count words), how much verbal direction, and whether to provide intermediate feedback as the agent proceeds.



Box 2 Figure i. The counting agent's world (top), with its starting and ending position (middle left and middle right) and (bottom left and right) its visual input at the beginning and the end of counting the objects.

Box 3: Outstanding Questions

What new developments in neural network architectures and learning algorithms will be necessary to address the challenge of creating a simulated agent that can pass the New York State Regent's exam in geometry?

What kinds of initial architectural constraints must be built in to the neural networks that will succeed in developing mathematical cognitive abilities, and how will the environment shape the emergent functional characteristics of the architecture?

What features of the natural environment support mathematics learning, and how do these features differ in different cultures?

What tasks and teaching policies are adopted by caregivers, peers and teachers in shaping the tasks and practices that shape learning?

What tools does the culture provide to support acquisition of mathematical ability? Stacking toys and shape sorters, tablets and markers, pencils and paper, rulers and compasses, slide-rules, calculators, and computers all can play a role in fostering the development of mathematical thinking.

What new insights into best practices for teaching and learning will a neural-network based approach afford?