Published in final edited form as:
*Behav Brain Sci*. 2011 June ; 34(3): 113–162. doi:10.1017/S0140525X10000919.

# The Origin of Concepts: A Precis

**Susan Carey**
Harvard University

The human conceptual repertoire is a unique phenomenon on earth, posing a formidable challenge to the disciplines of cognitive science. Alone among animals, humans can ponder the causes and cures of pancreatic cancer or global warming. How are we to account for the human capacity to create concepts such as *electron, cancer, infinity, galaxy*, and *wisdom*?

## 1. Components of a theory of conceptual development

As a matter of logic, a theory of conceptual development must have three components. First, it must characterize the innate representational repertoire—the representations that are the input to subsequent learning processes. Second, it must describe how the initial stock of representations differs from the adult conceptual system. Third, it must characterize the learning mechanisms that achieve the transformation of the initial into the final state.

*The Origin of Concepts* (TOOC) defends three theses. With respect to the initial state, contrary to historically important thinkers such as the British empiricists, Quine, and Piaget, as well as many contemporary scientists, the innate stock of primitives is not limited to sensory, perceptual or sensory-motor representations; rather, there are also innate conceptual representations. With respect to developmental change, contrary to continuity theorists such as Fodor, Pinker, Macnamara and others, conceptual development consists of episodes of qualitative change, resulting in systems of representation that are more powerful than and sometimes incommensurable with those from which they are built. With respect to a learning mechanism that achieves conceptual discontinuity, I offer Quinian bootstrapping.

## 2. Relations between theories of conceptual developmental and theories of concepts.

Obviously, our theory of conceptual development must mesh with our theory of concepts. Concepts are mental symbols, the units of thought. As with all mental representations, a theory of concepts must specify what it is that determines the content of any given mental symbol (i.e., what determines which concept it is, what determines the symbol's meaning). The theory must also specify how it is that concepts may function in thought—by virtue of what they combine to form propositions and beliefs, and how they function in inferential processes. TOOC assumes, and ultimately argues for, a dual factor theory of concepts. The two factors are sometimes called "wide" and "narrow" content. The contents of our mental representations are partly determined by causal connections between mental symbols, on the one hand, and the entities they refer to, on the other (wide content). To the extent this is so, all current psychological theories of concepts are partly on the wrong track—concepts are not prototypes, exemplar representations, nor theories of the entities they represent. The last chapter of TOOC reviews the arguments for information semantics--for wide content. However, contrary to philosophical views that deny that meanings are determined by what's in the head, TOOC argues that some aspects of inferential role are content determining (narrow content). The challenge for psychologists is saying what aspects of mental representations of entities we can think about partly determine the meaning of concepts of those entities, and which are simply what we believe about those entities (sometimes called

the project of distinguishing concepts from conceptions). One goal of TOOC is to explore how understanding of conceptual development requires a dual factor theory of concepts, and suggests how we might approach characterizing what aspects of conceptual role are content determining. Considerations of conceptual development constrain a theory of narrow content.

With the exception of developmental psychologists, cognitive scientists working on concepts have mostly abandoned the project of accounting for concept acquisition, for they have ignored the problem of characterizing and accounting for the features that enter into their learning models, often coding them with dummy variables.

This was not always so. In theorizing about concepts, the British empiricists made accounting for acquisition a central concern. They, like many modern thinkers, assumed that all concept learning begins with a primitive sensory or perceptual vocabulary. That project is doomed by the simple fact that it is impossible to express the meanings of most lexical items (e.g., “cause,” “good,” “seven,” “gold,” “dog…”) in terms of perceptual features. In response, some theorists posit a rich stock of developmental primitives, assuming that the definitional primitives that structure the adult conceptual repertoire and the developmental primitives over which hypothesis testing is carried out early in development are one and the same set. A moment’s reflection shows this assumption is also wrong. For example, the definition of *gold* within modern chemistry might be *element with atomic number 79*. Clearly, the primitives *element* and *atomic number* are not innate conceptual features. Or take the features that determine the prototype structure of bird concepts (flies, lays eggs, has wings, nests in trees, has a beak, sings, and so on…). Subjects provide distinctive values for such features when asked to list the features of birds, and overlap in terms of these same features predicts prototypicality within the category *bird*. That is, this feature space definitely underlies adult prototypicality structure. Yet these features are not innate primitives--many are no less abstract nor no less theory-laden than the concept *bird* itself. One of the goals of the TOOC is to characterize a learning process through which new primitives come into being.

## 3. The developmental primitives: core cognition

Explaining the human capacity for deep conceptual understanding begins with the observation that evolution provides developmental primitives, including the conceptual representations embedded in systems of core cognition, that are much richer than the sensori-motor representations that many hypothesize are the input to all learning. The first half of TOOC reviews the evidence for three domains of core cognition: the domain of middle-sized, middle-distant objects, including representations of causal and spatial relations among them (Chapters 2, 3 and 6), the domain of agents, including their goals, communicative interactions, attentional states, and causal potential (Chapters 5 and 6), and the domain of numbers, including parallel individuation, analog magnitude representations of the approximate cardinal values of sets, and set-based quantification (Chapters 4 and 7).

Core cognition resembles perception in many respects that distinguish it from other types of conceptual representations. These include the existence of innate perceptual input analyzers that identify the entities in core domains, a long evolutionary history, continuity throughout development, and iconic or analog format. The representations in core cognition differ from perceptual ones in having conceptual content.

### 3.1. Conceptual content

Two logically independent and empirically distinct properties of core cognition representations lead me to attribute them conceptual content. First, they cannot be reduced to

spatiotemporal or sensory vocabulary. One cannot capture concepts such as *goal, agent, object*, or *approximately 10* in terms of primitives such as locations, paths of motion, shapes, and colors. Second, they have a rich, central, conceptual role. The representations in core cognition are centrally accessible, represented in working memory models, and support action such as reaching—for example, infants make a working memory model of the individual crackers in each of two buckets and guide their choice of which bucket to crawl to from quantitative computations over those models. A variety of quantitative computations are defined over working memory model representations of sets of objects— preverbal infants can sum continuous quantities or compare models on the basis of 1-1 correspondence and categorically distinguish singletons from sets of multiple individuals. Moreover, young infants represent objects relative to the goals of agents, and infants' representations of physical causality are constrained by their conceptualization of the participants in a given interaction (as agents capable of self-generated motion or as inert objects). Thus, the conceptual status of the output of a given core cognition system is confirmed by its conceptual interrelations with the output of other core cognition systems.

In all other respects, the representations in core cognition resemble perceptual representations. Like representations of depth, the representations of objects, agents and number are the output of evolutionarily ancient, innate, modular input analyzers. Like the perceptual processes that compute depth, those that create representations of objects, agents and number continue to function continuously throughout the life span. And like representations of depth, their format is most likely iconic.

### 3.2 Dedicated input analyzers

A dedicated input analyzer computes representations of one kind of entity in the world and only that kind. All perceptual input analyzers are dedicated in this sense: the mechanism that computes depth from stereopsis does not compute color, pitch, or number.

Characterizing the mechanisms that identify the entities in systems of core cognition is important for several reasons, including that knowing how an input analyzer works bears on what aspects of the world it represents. Analysis of the input analyzers underlines the ways in which core cognition is perception like, and provides one source of evidence for the continuity of core cognition throughout development. For example, the primacy of spatiotemporal information in creating object representations and tracing object identity through time is one of the signatures that identifies infant object representations with those of mid-level object based attention and working memory in adults. Characterizations of the input analyzers bear on other theoretically important issues as well. For example, whether analog magnitude symbols are computed through a serial accumulation mechanism or by a parallel computation bears on whether it straightforwardly implements a counting algorithm, thus being likely to underlie learning to count (Chapters 4 and 7;) the evidence favors a parallel process. Or for another example, Chapter 5 presents evidence that infants use spatiotemporal descriptions of the motions and interactions among objects to assign agency, goals, and attentional states to them, and also that the static appearance of the entities in an event (e.g., presence of eyes and hands) also play a role in creating representations of agency. Such results raise the question of whether one of these sources of information is primary. For example, infants may initially identify agents through patterns of interaction, and may then learn what these agents look like. Alternatively, the innate face detectors infants have may serve the purpose of identifying agents, allowing them then to learn how agents typically interact. A third possibility is, like mother recognition in chicks, agency detection is such an important problem for human infants that evolution yielded two dedicated input analyzers to do the trick.

### 3.3. Innateness

The existence of dedicated input analyzers leaves open the question of their developmental history. What I mean for a representation to be innate is for the input analyzers that identify the represented entities to be the product of evolution, not the product of learning, and for at least some of its computational role to also be the product of evolution. For the most part, the evidence reviewed in TOOC for core cognition does not derive from experiments with neonates. Rather, the evidence for object representations comes from studies of infants 2 months of age or older, and that for representations of intentional agency by from infants 5 months of age or older. Two or 5 months is a lot of time for learning. Why believe that the representations tapped in these experiments are the output of *innate* input analyzers, and why believe that the demonstrated inferential role that provides evidence for the content of the representations is unlearned? I discuss this question in each case study, appealing to four types of arguments.

First, success at some task provides support for some target representational capacity needed to carry it out, whereas failure is not necessarily good evidence that the target capacity is lacking. Some other representational capacity, independent of the target one, may be needed for the task and may not yet be available (not yet learned or not yet matured). TOOC provides several worked out examples of successful appeals to performance limitations masking putatively innate competences. For example, the A/not B error between ages 7 to 12 months is at least in part explained by appeal to immature executive function. Or for another example, that it is not until 2 months of age that infants create representations of a complete rod partially hidden behind a barrier when they are shown the protruding ends undergoing common motion is at least partly explained by infants' failure, below 2 months of age, to notice the common motion across the barrier, so they lack the critical input to the putatively innate computation.

Secondly, that a given representational capacity may be innate in humans, in spite of not being observed until some months after birth, is suggested by evidence that it is manifest in neonates of other species. Examples offered were depth perception, which emerges without opportunities for learning in neonate goats and neonate rats, and object representations, which are observed in neonate chicks. This line of evidence is obviously indirect, providing only an existence proof that evolution can build input analyzers that create representations with the content in question.

Thirdly, the simultaneous emergence of different aspects of a whole system also provides indirect evidence for the innateness of the input analyzers and computational machinery that constitutes core cognition. As soon as infants can be shown to form representations of complete objects, only parts of which had been visible behind barriers, they also can be shown to use evidence of spatiotemporal discontinuity to infer that two numerically distinct objects are involved in an event, and also to represent object motion as constrained by solidity (Chapters 2 and 3). Similarly, different aspects of intentional attribution emerge together; representing an entity as capable of attention increases the likelihood of representing its action as goal directed, and vice versa (Chapter 5). If the generalizations that underlie infants' behavior are learned from statistical analyses of the input (represented in terms of spatiotemporal and perceptual primitives), it is a mystery why all of the interrelated constraints implicated in the core cognition proposals emerge at once. Infants have vastly different amounts of input relevant to different statistical generalizations over perceptual primitives. Relative to the thousands of times they have seen objects disappear behind barriers, two-month-old infants have probably never seen rods placed into cylinders, and rarely seen solid objects placed into containers. Yet the interpretation of both types of events in terms of the constraints on object motion that are part of core cognition emerge together,

at 2 months of age. Statistical learning of perceptual regularities would be expected to be piecemeal, not integrated.

Finally, learnability considerations also argue that the representations in core cognition are the output of innate input analyzers. If the capacity to represent individuated objects, numbers, and agents are learned, built out of perceptual and spatiotemporal primitives, then there must be some learning mechanism capable of creating representations with conceptual content that transcend the perceptual vocabulary. In the second half of TOOC, I offer Quinian bootstrapping as a mechanism that could, in principle, do the trick, but this type of learning process requires explicit external symbols (words, mathematical symbols), and these are not available to young babies. Associative learning mechanisms could certainly come to represent regularities in the input, such as that if a bounded stimulus disappeared through deletion of the forward boundary behind another bounded stimulus there is a high probability that a bounded stimulus resembling that that disappeared will appear by accretion of the rear boundary from the other side of the constantly visable bounded surface. But these generalizations would not be formulated in terms of the concept *object*. There is no proposal I know for a learning mechanism available to non-linguistic creatures that can create representations of objects, number, agency, or causality from perceptual primitives.

### 3.4 Iconic format

A full characterization of any mental representation must specify its format as well as its content and conceptual role. What are the mental symbols like? How are they instantiated in the brain? I intend the distinction between iconic and non-iconic formats to be the same distinction that was at stake in the historical debates on the format of representation underlying mental imagery. Iconic representations are analog; roughly, the parts of the representation correspond to the parts of the entities represented.

We know little about the format of most mental representations. Of the core cognition systems discussed in TOOC the question of format is clearest for number representations, so my discussion of format was concentrated there (Chapter 4). The very name of analog magnitude representations stakes a claim for their format. Analog representations of number represent as would a number line—the representation of 2 (—) is a quantity that is smaller than and is contained in the representation for 3 (—). We do not know how these analog representations are actually instantiated in the brain—larger quantities could be represented by more neurons firing or by faster firing of a fixed populations of neurons, for example. Many plausible models have been proposed (see Chapter 4). However analog magnitude representations are instantiated in the brain, their psychophysical signatures strongly suggest this type of representational scheme. That discrimination satisfies Weber's law (is a function of the ratio of set sizes) suggests that number representations work like represntations of length, time, area, brightness and loundness. All proposals for how all of these continuous dimensions are represented also deploy analog magnitudes.

TOOC speculates that all of core cognition is likely to be represented in iconic format. Consider the working memory models constitute the parallel individuation system of object representations. The fact that these representations are subject to the set-size limit of parallel individuation implicates a representational schema in which each individual in the world is represented by a symbol in working memory. This fact does not constrain the format of these symbols. A working memory model for two boxes of different front surface areas, for instance, could be consist of image-like representations of the objects (□□), or they could be abstract symbols (*object(3 square inches), object(4 square inches)*). These models must include some representation of size bound to each symbol for each object, because the total volume or total surface area of the objects in a small set is computable from the working memory representations of the sets. The most plausible model for how this is done

implicates iconic representations of the objects, with size imagistically represented, as well as shape, color, and other perceptual properties bound to the symbols iconically. The iconic alternative laid out in Chapter 4 explains the set size limits on performance even when continuous variables are driving the response.

I have several other reasons for suspecting that the representations in core cognition are iconic. Iconic format is consistent with (through not required by) the ways in which the representations in core cognition are perception like. Second, just as static images may be iconic or symbolic, so too may representations of whole events. If infants represent events in iconic format, like a movie that can be replayed, this could help make sense of the apparently retrospective nature of the representations that underlie many violation of expectancy looking time experiments (Chapters 2-6). Finally, that core cognition may be represented in terms of iconic symbols, with some of its content captured in encapsulated computations defined over these symbols, may help to make sense of the extreme lags between understanding manifest in infant looking time studies and that manifest only much later in tasks that require explicit linguistic representations (see Chapters 3, 5, 8-12). The guess that the format of *all* core cognition is iconic is just that—a guess—but the considerations just reviewed lead me to favor this hypothesis.

### 3.5 Constant through the life span

One might think that any representations important enough that evolution created dedicated perceptual input devices to detect specific classes of entities in the world, and important enough that evolution built specialized inferential machinery for thinking about those entities, should be useful for adults as well as children. However, this first thought is not necessarily correct. Some innate representational systems serve only to get development off the ground. The learning processes (there are two) that support chicks recognizing their mother, for example, operate only in the first days of life, and their neural substrate actually atrophies when their work is done. Also, given that some of the constraints built into core knowledge representations are overturned in the course of explicit theory building, it is at least possible that the core cognition system itself might be overridden.

Thus, it is most definitely an empirical question whether core cognition is constant throughout the life span. That the core cognition systems described in TOOC are is supported by the same signatures of processing in adulthood and infancy. Under conditions where core cognition is isolated from other conceptual resources, adults display the same limits on computations, and the same modular input analyzers, as do infants. Continuity through the life span is an important property of core cognition for several reasons. We seek an account of cognitive architecture that carves the mind into meaningful sub-systems; and *most* conceptual representations are *not* continuous throughout development. Core cognition is one very distinctive part of the human mind: no other system of conceptual representations shares its suite of characteristics.

### 3.6. A Dual Factor Theory of the Concepts within Core Cognition

Information semantics (wide content) has pride of place in a theory of core cognition. This is because the representations within core cognition are the output of innate perceptual input analyzers. These input analyzers come into being through natural selection, a process that, in this case, explains how the extension of the concepts within core cognition may be determined by causal connections between entities in their domains (objects, agents, goals, cardinal values of sets, and the like) and the mental symbols that represent them. Moreover, since there are aspects of innate conceptual role that remain constant throughout development, these unproblematically specify the narrow content of representations within

core cognition. Dual factor theory straightforwardly applies to the representations in core cognition.

## 4. Beyond core cognition: central innate representations.

### 4.1 Representations of *cause*

The existence of innate conceptual representations embedded within systems of core cognition does not preclude other innate conceptual representations as well, including non-domain specific central ones. Chapter 6 takes the concept *cause* as a case study, and contrasts Michotte's proposal that innate causal representations are part of core object cognition with the proposal that there may be innate central representations of causation. On Michotte's proposal, the earliest causal representations should be sensitive only to spatiotemporal relations among events, and should be limited to reasoning about causes of object motion. An impressive body of empirical data establishes that by 6 months of age, infants represent Michottian motions events (launching, entraining and expulsion) causally. Nonetheless, Chapter 6 rejected Michotte's proposal on the grounds that causal cognition integrates across different domains of core cognition (object representations and agent representations), encompassing state changes as well as motion events, from as early in development as we have evidence for causal representations at all.

Innate central causal representations could come in either of two quite different varieties. There may be innate central processes that compute causal relations from patterns of statistical dependence among events, with no constraints on the kinds of events. Or there may be specific aspects of causality that are part of distinct core cognition systems (e.g., Michottian contact causality within the domain of core object cognition and intentional causality within the domain of agent cognition) and these may be centrally integrated innately. These possibilities are not mutually exclusive; both types of central integration of causal representations could be part of infants' innate endowment.

### 4.2. Public symbols, logical and linguistic capacity

OC is mostly silent with respect to two important aspects of conceptual development. I assume that domain specific learning mechanisms, jointly comprising a language acquisition device (LAD), make possible language acquisition, but I have made no effort to summarize the current state of the art in characterizing the LAD. Whatever its nature, the LAD is another way innate cognitive architecture goes beyond core cognition, for the symbols in language are not iconic. Second, I have said almost nothing about the logical capacities humans and other animals are endowed with, although these are independent of core knowledge and I help myself to various of them in my bootstrapping proposals. These are topics for other books.

Language acquisition and conceptual development are intimately related. The representations in core cognition support language learning, providing some of the meanings that languages express. Chapter 7 considered how prelinguistic set-based quantification supports the learning of natural language quantifiers, and prelinguistic representations of individuals support the learning of terms that express sortals. But because my concern is the origin of concepts, I focused mainly on the complementary relations between language learning on conceptual development. Language learning makes representations more salient or efficiently deployed (Chapter 7; so-called weak effects of language learning on thought), and plays a role in conceptual discontinuities (strong effects of language learning on thought.)

Chapter 7 reviewed two cases in which very early language learning affects non-linguistic representations. First, learning, or even just hearing, labels for objects influences the

establishing/deploying of sortal concepts. Second, mastery of explicit linguistic singular/plural morphology plays a role in deploying this quantificational distinction in non-linguistic representations of sets. Although TOOC argues that these are most likely weak effects of language learning on thought, "weak" does not mean the same thing as "uninteresting" or "unimportant." Creating representations whose format is non-iconic paves the way for integrating the concepts in core cognition with the rest of language.

Furthermore, most of the second half of the book concerns how language learning also shapes thought in the strongest possible way. Language learning plays a role in creating new representational resources that include concepts previously unrepresentatable.

## 5. Discontinuity—the descriptive problem

Discontinuity in conceptual development arises at two different levels of abstraction. In terms of basic cognitive architecture, core cognition differs qualitatively from explicit linguistically encoded knowledge. Consider the concepts *planet* or *germ*. These concepts are not the output of innate input analyzers, and so are neither innate nor causally connected to the entities they represent in the same way as are the concepts in core cognition. They are not evolutionary ancient. Unlike core cognition representations, their format is certainly not iconic, and they are not embedded in systems of representation that are constant over development. Explicit conceptual representations can be, and often are, overturned in the course of conceptual development. Thus, in terms of general cognitive architecture, explicit, verbally represented, intuitive theories are qualitatively different from, and hence discontinuous with systems of core cognition.

Conceptual discontinuities are found at a more specific level as well—discontinuities within particular content domains. Establishing conceptual discontinuity at this level requires specifying the qualitative differences between two successive conceptual systems (CS1 and CS2). In some of the case studies in TOOC, new representational resources are constructed with more expressive power than those from which they are built. In other cases, theories are constructed whose concepts are incommensurable with those from which they are built. Both types of discontinuity (increased expressive power, incommensurablity) involve *systems* of concepts and inferences, and so evidence for discontinuity must include evidence of within-child consistency over a wide range of probes of the underlying representational capacity. Also, discontinuity implies mastery of CS2 should be difficult, and there should be initial assimilation of input couched in the language of CS2 in terms of the concepts of CS1.

## 6. Discontinuities in the sense of increased expressive power; mathematical concepts.

### 6.1 Natural Number.

Core cognition contains two systems of representation with numerical content: parallel individuation of small sets of entities in working memory models, and analog magnitude representations of number. Within, LAD, a third innate system of representation with numerical content supports the learning of natural language quantifiers. These are the CS1s. CS2, the first explicit representational system that represents the positive integers, is the verbal numeral list embedded in a count routine. Deployed in accordance with the counting principles articulated by Gelman and Gallistel, the verbal numerals implement the successor function, at least with respect to the child's finite count list. For any numeral that represents cardinal value $n$, the next numeral in the list represents $n + 1$.

CS2 is qualitatively different from each of the CS1s because none of the CS1s has the capacity to represent the integers. Parallel individuation includes no summary symbols for

number at all, and has an upper limit of 3 or 4 on the size of sets it represents. The set-based quantificational machinery of natural language includes summary symbols for quantity (*plural, some, all*), and importantly contains a symbol with content that overlaps considerably with that of "one" (namely, the singular determiner, "a"), but the singular determiner is not embedded within a system of arithmetical computations. Also, natural language set-based quantification has an upper limit on the sets sizes that are quantified with respect to exact cardinal values (*singular, dual, trial*). Analog magnitude representations include summary symbols for quantity that are embedded within a system of arithmetical computations, but they represent only approximate cardinal values; there is no representation of exactly 1, and therefore no representation of + 1. Analog magnitude representations cannot even resolve the distinction between 10 and 11 (or any two successive integers beyond its discrimination capacity), and so cannot express the successor function. Thus, none of the CS1s can represent *10*, let alone *342,689,455*.

This analysis makes precise the senses in which the verbal numeral list (CS2) is qualitatively different from those representations that precede it: it has a totally different format (verbal numerals embedded in a count routine) and more expressive power than any of the CS1s that are its developmental sources.

As required by CS2's being qualitatively different from each of the CS1s that contain symbols with numerical content, it is indeed difficult to learn. American middle-class children learn to recite the count list and to carry out the count routine in response to the probe "how many," shortly after their second birthday. They do not learn how counting represents number for another 1 ½ or 2 years. Young two-year-olds first assign a cardinal meaning to "one," treating other numerals as equivalent plural markers that contrast in meaning with "one." Some 7 to 9 months later they assign cardinal meaning to "two," but still take all other numerals to mean essential "some," contrasting only with "one" and "two." They then work out the cardinal meaning of "three" and then of "four." This protracted period of development is called the "subset"-knower stage, for children have worked out cardinal meanings for only a subset of the numerals in their count list.

Many different tasks, which make totally different information processing demands on the child, confirm that subset-knowers differ qualitatively from children who have worked out how counting represents number. Subset-knowers cannot create sets of sizes specified by their unknown numerals, cannot estimate the cardinal values of sets outside their known numeral range, do not know what set-size is reached if 1 individual is added to a set labeled with a numeral outside their known numeral range, and so on. Children who succeed on one of thee tasks succeed on all of them. Furthermore, a child diagnosed as a "one"-knower on one task is also a "one"-knower on all of the others, ditto for "two"-knowers, "three"-knowers and "four"-knowers. The patterns of judgments across all of these tasks suggest that parallel individuation and the set-based quantification of natural language underlie the numerical meanings subset-knowers construct for numeral words.

In sum, the construction of the numeral list representation is a paradigm example of developmental discontinuity. How CS2 transcends CS1 is precisely characterized, CS2 is difficult to learn, adult language expressing CS2 is assimilated to CS1, and children's performance on a wide variety of tasks consistently reflect either CS1 or CS2.

### 6.2 Rational Number.

Chapter 9 presents another parade case of developmental discontinuity within mathematical representations. In CS1, the count list representation of the positive integers, *number* means natural number. Early arithmetic instruction depends upon and further entrenches this representational system, representing addition and subtraction as counting up and counting

down, and modeling multiplication as repeated addition. In CS2 *number* means any point on a number line that can be expressed x/y, where x and y are integers. In CS2, rather than it being the case that integers are the only numbers, there are an infinity of numbers between any two integers. The question of the next number after n (where n might be an integer or not) no longer has an answer. Thus, CS2 has more expressive power than CS1 (there are all of these extra numbers), and numbers are related to each other differently in the two systems. The new relation in CS2 is division. Division cannot be represented in terms of the resources of CS1, which model only addition, subtraction and multiplication of integers. CS2's division cannot be represented as repeated subtraction of integers.

CS2 is extremely difficult for children to learn. One half of college bound high school students taking the SAT exams do not understand fractions and decimals. Furthermore, explicit instruction concerning rational number is initially assimilated to CS1, and children are consistent over a wide range of probes as to how they conceptualize number. Whether children can properly order fractions and decimals, how they justify their ordering, how they explain the role of each numeral in a fraction expressed "x/y", whether they agree there are numbers between 0 and 1 and whether they believe that repeated division by 2 will ever yield 0 are all interrelated. What the child does on one of these tasks predicts what he/she will do on all of the others. CS1 and CS2 are each coherent conceptual systems, qualitatively different from each other.

## 7. Discontinuities in the sense of local incommensurability: natural kind concepts

Conceptual discontinuity is not only a matter of increased expressive power. Sometimes, two successive conceptual systems are qualitatively different because they are locally incommensurable and thus not mutually translatable. One cannot express the beliefs that articulate CS2 in the concepts of CS1 and vice versa.

Incommensurability arises when episodes of conceptual development have required conceptual change. Conceptual changes are of several kinds, including differentiations such that the undifferentiated concept in CS1 plays no role in CS2, and is even incoherent from the point of view of CS2, coalescences in which ontologically distinct entities from the point of view of CS1 are subsumed under a single concept in CS2, and changes in conceptual type and in content-determining conceptual cores.

The analysis of local incommensurability in TOOC illustrates the fruits of what Nancy Nersessian calls "cognitive historical analysis," in which philosophers and historians of science join forces with cognitive scientists to understand knowledge acquisition both in the history of science and over individual ontogenesis. The theory-theory of cognitive development presupposes that the same questions can be asked of episodes of knowledge acquisition in individual children and historical theory changes, in spite of the manifest differences between scientists and children, and that sometimes these questions receive the same answers in the two cases. Examples are: what is an "undifferentiated concept," what counts as evidence for lack of conceptual differentiation, what distinguishes episodes of conceptual development that merely involve belief revision from those involving conceptual change?

Conceptual change occurs when sets of concepts that are interdefined are acquired together, en suite, with content determining interconnections that differ from those in CS1 and with new concepts emerging that are not representable in CS1. Chapter 10 sketches an historical example of conceptual change between the source-recipient and caloric theories of thermal phenomena, focusing on the differentiation of the concepts *heat* and *temperature*. The

developmental example juxtaposed to this involves incommensurable intuitive theories of the physical world, focusing of the differentiation of the concepts *physical* and *material* and the concepts *weight* and *density*.

Chapter 10 describes many phenomena that suggest that children's concepts of the physical world may be incommensurable with ours: their confidence that a small piece of styrofoam weighs 0 grams, non-conservation of amount of matter and of weight, the claim that dreams are made of air, that shadows exist in the dark but we just can't see them. At the heart of establishing local incommensurability is characterizing two successive physical theories, providing evidence that each is a theory children actually hold, and, of course, displaying the incommensurability. Chapter 10 characterizes an initial theory (CS1), in which an undifferentiated concept *weight/density* functions coherently. A translator's gloss is provided, sketching the central concept *degree of heaviness* akin to the Florentine Experimenter's *degree of heat*, which was analogously undifferentiated between *heat* and *temperature*. A sketch of CS1's concept *physically real/substantial*, the concept closest to CS2's *material*, was also part of the translator's gloss, as was a sketch of the undifferentiated concept *air/nothing*. CS1's undifferentiated concepts cannot be expressed in terms of any conceptual system that differentiates them; they are incoherent from the point of view of CS2.

Chapter 10 also characterizes conceptual changes other than differentiation. It documents ancestor concepts in CS1 that represent kinds as ontologically distinct which CS2 unites under a single concept. An example is CS2's *matter*, uniting what are vastly different kinds in CS1 (*object, liquid, air*). Ancestor concepts in CS1 also differ from their descendents in CS2 in type and features taken to be essential. The essential features of CS1's undifferentiated concept *matter/physically real* are perceptual access and causal interaction with other external physical entities. The essential features of the CS2's *matter* are weight and occupying space. An interconnected change occurs within the concept *degree of heaviness*. In CS1, degree of heaviness is a property of *some* material/physically real entities, such as a large piece of Styrofoam but not a small piece. In CS2, weight is taken to be an essential feature of all material entities, a property that provides an extensive measure of amount of matter. The local incommensurability between CS1 and CS2 derives from the simultaneous adjusting these concepts to each other. Differentiations implicating incommensurability never occur independently of simultaneous coalescences, nor of changes of the causally deepest properties known of each of a system of interrelated concepts.

If this analysis is correct, CS2 should be difficult to learn, and indeed it is. Although the target of science instruction, a large proportion of secondary school students fail to undergo the conceptual change. Finally, there is striking within-child consistency across the many disparate tasks that diagnose CS1 and CS2: sorting entities as matter/non-matter, representing matter, weight, and volume as continuous extensive variables, modeling weight, density and volume of a set of objects, ordering objects with respect to weight, density and volume, and measuring weight, density and volume.

Pondering children's responses on the tasks probing these concepts is what allows the reader to come to represent CS1. Constructing a set of coherent concepts that yield the same judgments as those of children with CS1 is a bootstrapping process. Aided by the translator's gloss, the reader must create a conceptual system in which *degree of heaviness* functions coherently.

## 8. Quinian Bootstrapping

Ultimately learning requires adjusting expectations, representations, and actions to data. Abstractly, all of these learning mechanisms are variants of hypothesis testing algorithms. The representations most consistent with the available data are strengthened; those hypotheses are accepted. However, in cases of developmental discontinuity, the learner does not initially have the representational resources to state the hypotheses that will be tested, to represent the variables that could be associated or could be input to a Bayesian learning algorithm. Quinian bootstrapping is one learning process that can create new representational machinery, new concepts that articulate hypotheses previously unstatable.

In Quinian bootstrapping episodes, mental symbols are established that correspond to newly coined or newly learned explicit symbols. These are initially placeholders, getting whatever meaning they have from their interrelations with other explicit symbols. As is true of all word learning, newly learned symbols must necessarily be initially interpreted in terms of concepts already available. But at the onset of a bootstrapping episode, these interpretations are only partial—the learner (child or scientist) does not yet have the capacity to formulate the concepts the symbols will come to express.

The bootstrapping process involves modeling the phenomena in the domain, represented in terms of whatever concepts the child or scientist has available, in terms of the set of interrelated symbols in the placeholder structure. Both structures provide constraints, some only implicit and instantiated in the computations defined over the representations. These constraints are respected as much as possible in the course of the modeling activities, which include analogy construction and monitoring, limiting case analyses, thought experiments, and inductive inference.

### 8.1. Bootstrapping representations of natural number

TOOC draws on Quinian bootstrapping to explain all the developmental discontinuities sketched above. In the case of the construction of the numeral list representation of the integers, the memorized count list is the placeholder structure. Its initial meaning is exhausted by the relation among the external symbols: they are stably ordered. “One, two, three, four…” initially has no more meaning for the child than “a, b, c, d…” The details of the subset-knower period suggest that the resources of parallel individuation, enriched by the machinery of linguistic set-based quantification, provide the partial meanings children assign to the placeholder structures that get the bootstrapping process off the ground. The meaning of the word “one” could be subserved by a mental model of a set of a single individual {i}, along with a procedure that determines that the word “one” can be applied to any set that can be put in 1-1 correspondence with this model. Similarly “two” is mapped onto a long term memory model of a set of two individuals {j k}, along with a procedure that determines that the word “two” can be applied to any set that can be put in 1-1 correspondence with this model. And so on for “three” and “four.” This proposal requires no mental machinery not shown to be in the repertoire of infants—parallel individuation, the capacity to compare models on the basis of 1-1 correspondence, and the set-based quantificational machinery that underlies the singular/plural distinction and makes possible the representation of dual and trial markers. The work of the subset-knower period of numeral learning, which extends in English-learners between ages 2:0 and 3:6 or so, is the creation of the long term memory models and computations for applying them that constitute the meanings of the first numerals the child assigns numerical meaning to.

Once these meanings are in place, and the child has independently memorized the placeholder count list and the counting routine, the bootstrapping proceeds as follows: The child notices the identity between the singular, dual, trial, and quadral markers and the first

four words in the count list. The child must try to align these two independent structures. The critical analogy is between order on the list and order in a series of sets related by *additional individual*. This analogy supports the induction that any two successive numerals will refer to sets such that the numeral farther in the list picks out a set that is 1 greater than that earlier in the list.

This proposal illustrates all of the components of bootstrapping processes: placeholder structures whose meaning is provided by relations among external symbols, partial interpretations in terms of available conceptual structures, modeling processes (in this case analogy), and an inductive leap. The greater representational power of the numeral list than that of any of the systems of core cognition from which it is built derives from combining distinct representational resources—a serially ordered list, set-based quantification (which gives the child singular, dual, trial, and quadral markers, as well as other quantifiers), and the numerical content of parallel individuation (which is largely embodied in the computations carried out over sets represented in memory models with one symbol for each individual in the set). The child creates symbols that express information that previously existed only as constraints on computations. Numerical content does not come from nowhere, but the process does not consist of defining "seven" in terms of mental symbols available to infants.

### 8.2. Bootstrapping in the construction of explicit scientific theories

Historians and philosophers of science, as well as cognitive scientists, working with daily records of scientists' work, have characterized the role of Quinian bootstrapping in scientific innovation. Chapter 11 draws out some of the lessons from case studies of Kepler, Darwin and Maxwell.

In all three of these historical cases, the bootstrapping process was initiated by the discovery of a new domain of phenomena that became the target of explanatory theorizing. Necessarily, the phenomena were initially represented in terms of the theories available at the outset of the process, often with concepts that were neutral between those theories and those that replaced them. Incommensurability is always local; much remains constant across episodes of conceptual change. For Kepler, the phenomena were the laws of planetary motion; for Darwin, they were the variability of closely related species and the exquisite adaptation to local environmental constraints; for Maxwell, they were the electromagnetic effects discovered by Faraday and others.

In all three of these cases the scientists created an explanatory structure that was incommensurable with any available at the outset. The process of construction involved positing placeholder structures and involved modeling processes which aligned the placeholders with the new phenomena. In all three cases, this process took years. For Kepler, the hypothesis that the sun was somehow causing the motion of the planets was a placeholder until the analogies with light and magnetism allowed him to formulate *vis motrix*. For Darwin, the source analogies were artificial selection and Mathus' analysis of the implications of a population explosion for the earth's capacity to sustain human beings. For Maxwell, a much more elaborate placeholder structure was given by the mathematics of Newtonian forces in a fluid medium. These placeholders were formulated in external symbols—natural language, mathematical language, and diagrams.

Of course, the source of these placeholder structures in children's bootstrapping is importantly different from that of scientists. The scientists posited them as tentative ideas worth exploring, whereas children acquire them from adults, in the context of language learning or science education. This difference is one reason why metaconceptually aware hypothesis formation and testing is likely to be important in historical cases of conceptual

change. Still, many aspects of the bootstrapping process are the same whether the learner is a child or a sophisticated adult scientist. Both scientists and children draw on explicit symbolic representations to formulate placeholder structures and on modeling devices such as analogy, thought experiments, limiting case analyses, and inductive inference to infuse the placeholder structures with meaning.

### 8.3. Bootstrapping processes underlying conceptual change in childhood

Historically, mappings between mathematical structures and physical ones have repeatedly driven both mathematical development and theory change. In the course of creating the theory of electromagnetic fields, Maxwell invented the mathematics of quantum mechanics and relativity theory. Another salient example is Newton's dual advances in the calculus and in physics.

So too in childhood. Constructing mappings between mathematical and physical representations plays an essential role both in the creation of concepts of rational number and the creation of theory of the physical world in which weight is differentiated from density. These two conceptual changes constrain each other. The child's progress in conceptualizing the physical world exquisitely predicts understanding of rational number and vice-versa. Children whose concept of number is restricted to positive integers have not yet constructed a continuous theory of matter nor a concept of weight as an extensive variable, whereas children who understand that number is infinitely divisible have done both.

Carol Smith's bootstrapping curriculum provides insight into the processes through which *material* becomes differentiated from *physically real* and *weight* from *density*. Although developed independently. Smith's curriculum draws on *all* of the components of the bootstrapping process that Nersessian details in her analysis of Maxwell. First, Smith engages students in explaining new phenomena, ones that can be represented as empirical generalizations stated in terms of concepts they already have. These include the proportionality of scale readings to overall size (given constant substance), explaining how different sized entities can weigh the same, predicting which entities will float in which liquids, and sorting entities on the basis of whether they are material or immaterial, focusing particularly on the ontological status of gases. She then engages students in several cycles of analogical mappings between the physical world and the mathematics of extensive and intensive variables, ratios and fractions. She begins with modeling the extensive quantities of weight and volume with the additive and multiplicative structures underlying integer representations. The curriculum then moves to calculating the weight and volume of very small entities, using division. These activities are supported by thought experiments (which are themselves modeling devices) that challenge the child's initial concept *weight* as *felt weight/density*, leading them into a contradiction between their claim that a single grain of rice weighs 0 grams, and the obvious fact that 50 grains of rice have a measurable weight. Measuring the weight of a fingerprint and a signature with an analytical balance makes salient the limits of sensitivity of a given measurement device, and further supports conceptualizing weight as an extensive variable that is a function of amount of matter.

To complete the differentiation of *weight* from *density* Smith makes use of visual models that represent the mathematics of extensive and intensive quantities. The visual models consist of boxes of a constant size, and numbers of dots distributed equally throughout the boxes. Numbers of dots and numbers of boxes are the extensive variables, numbers of dots per box the intensive variable. Students first explore the properties of these objects in themselves, discovering that one can derive the value of any one of these variables knowing the values of the other two, and exploring the mathematical expression of these relations: dots per box = number of dots divided by number of boxes. The curriculum then moves to

using these visual objects to model physical entities, with number of boxes representing volume and number of dots representing weight. Density (in the sense of weight/volume) is visually represented in this model as dots/box, and the models make clear how it is that two objects of the same size might weigh different amounts--because they are made of materials with different densities, why weight is proportional to volume given a single material, and so on. The models are also used to represent liquids, and students discover the relevant variables for predicting when one object will float in a given liquid and what proportion of the object will be submerged. This activity is particularly satisfying for students, because at the outset of the curriculum, with their undifferentiated *weight/density* concept, they cannot formulate a generalization about which things will sink and which will float. Differentiating *weight* from *density* in the context of these modeling activities completes the construction of an extensive concept *weight* begun in the first part of the curriculum.

The formula D = W/V (density equals weight divided by volume) is a placeholder structure at the beginning of the bootstrapping process. The child has no distinct concepts *weight* and *density* to interpret the proposition. As in all cases of bootstrapping, the representation of placeholder structures makes use of the combinatorial properties of language. Density here is a straightforward complex concept, defined in terms of a relation between weight and volume (division), and the child (if division is understood) can understand this sentence as: something equals something else divided by something (most children also have no concept of volume at this point in development). The dots per box model is also a placeholder structure at the beginning of the bootstrapping process, a way of visualizing the relations between an intensive variable and two extensive variables related by division, and thus provides a model that allows the child to think with, just at Maxwell's models allowed him to think with the mathematics of Newtonian mechanics as he tried to model Faraday's phenomena. At the outset of the process the child has no distinct concepts *weight* and *density* to map to number of dots and number of dots/box, respectively.

Although straightforward conceptual combination plays a role in these learning episodes (in the formulation of the placeholder structures), the heart of Quinian bootstrapping is the process of providing meaning for the placeholder symbols. At the outset *weight* is interpreted in terms of the child's undifferentiated concept *degree of heaviness* (see Chapter 10) and *density* has no meaning whatsoever for the child. It is in terms of the undifferentiated concept *degree of heaviness* that the child represents the empirical generalizations that constitute the phenomena he or she is attempting to model. The placeholder structure introduces new mental symbols (*weight* and *density*). The modeling processes, the thought experiments and analogical mapping processes, provides content for them. The modeling process, using multiple iterations of mappings between the mathematical structures and the physical phenomena, makes explicit in a common representation what was only implicit in one or the other representational systems being adjusted to each other during the mapping.

That bootstrapping processes with the same structure play a role in conceptual change both among adult scientists and young children is another fruit of cognitive historical analysis. As I have repeatedly mentioned, a point bearing further repetition, this does not belie important differences between adult scientists engaged with metaconceptual awareness in explicit theory construction and young children. Without denying these differences, chapters 8 – 12 illustrate what the theory-theory of conceptual development buys us. By isolating questions that receive the same answers in each case, we can study conceptual discontinuities and the learning mechanisms that underlie them, bringing hard won lessons from each literature to bear on the other.

## 10, Implications for a theory of concepts

The first twelve chapters of the book stand alone in painting a picture of how conceptual development is possible. The thirteenth steps back and explores the implications of what has been learned for a theory of concepts. I detail a range of phenomena that prima facie we might think a theory of concepts should explain. A theory of concepts should contribute to our understanding of the human capacities for categorization and productive thought (conceptual combination and inference). It must explain how mental symbols refer, how they support epistemic warrant, and how, at least in principle, they may be acquired. And it must distinguish between our concepts of entities in the world and our beliefs and knowledge of them.

As many commentators have noted, the empiricists' theory (sometimes called the "classical view") was unmatched in its scope, offering machinery that met all of the above desiderata. Of particular importance to me, the actual empiricists' classical view put explaining concept acquisition front and center. Chapter 13 summarizes the empiricists' theory, and sketches why both psychologists and philosophers abandoned it as hopeless.

TOOC is an extended argument against the classical view of concept acquisition. The primitives from which our concepts are built are not solely sensory, and the processes of building them are not exhausted by assembling definitions (or even representations of prototypes) in terms of the primitive base. For the most part though, the failure to account for acquisition was not the reason most psychologists abandoned the empiricists' theory. Rather, the culprit was its failure to account for categorization data. However, many considerations weigh against making categorization processes the central explanandum for a theory of concepts. Writers as disparate as Greg Murphy and W. V. O. Quine have argued that categorization is a holistic process, drawing on everything we know about the entities that fall under an concept, and I agree. But this is precisely why experiments on categorization decisions cannot be a pipeline to our analysis of concepts. We cannot isolate the contribution of conceptual content from the beliefs we hold about the entities that fall under a concept (concepts from conceptions) in determining the data from categorization experiments.

In my view, more telling arguments against the classical view derive from the philosophical literature on information semantics and wide content, and from the literature on psychological essentialism that was inspired by it. Kripke's analysis of the processes that fix the reference of proper names and Kripke's and Putnam's extension of this analysis to natural kind terms convince me the reference is not determined solely by what's in the head. This work undermines not only the classical view of concepts, but also prototype/exemplar theories and a purely internalist theory-theory of concepts. The phenomena in support of psychological essentialism point to the same conclusion. We can and do deploy our concepts in the absence of any knowledge of the entities that fall under them that would allow us to pick out those entities, and under the assumption that everything we know about them is revisable. At least in the case of natural kind concepts, we assume that what determines their extension is a matter for science to discover, not for us to stipulate as a matter of definition, nor for our current prototypes or theories to determine, even probabilistically. In my view, these considerations force a role for wide content in a theory of concepts.

One theory of information semantics, Fodor's, also places concept acquisition front and center. Fodor's view, at least at one time, led him to a radical concept nativism, according to which all concepts the grain of single lexical items are innate. OC is also an extended argument against radical concept nativism. Chapter 13 argues against a pure information theory; we cannot ignore what is inside the head. Conceptual role has potentially three

different roles to play in a theory of concepts. First, concepts underlie thought, inference, and planning, and whatever we take concepts to be must be compatible with these functions. Second, on at least some approaches to information semantics, recognition processes are part of the causal connection between entities in the world and the mental symbols that represent them. Although the mechanisms that support categorization decisions may not determine content in the way that traditional theories envisioned, they may have a place in a theory of how reference is determined. A dual factor theory requires more though—it requires showing that some aspects of conceptual role are genuinely content determining. That is, it requires a commitment to narrow content. Chapter 13 details many arguments for narrow content. Some of the desiderata for a theory of concepts, including accounting for conceptual productivity and accounting for incommensurability, require appeals to conceptual role. Conceptual role exhausts the content of logical connectives, and surely has pride of place in content determination of mathematical symbols. Furthermore, Ned Block showed that even accepting the Kripke/Putnam arguments for wide content, conceptual role still has a place in the determination of the extension natural kind concepts. It is a fact about our psychology that a given concept is a natural kind concept, i.e., is assumed under psychological essentialism. That is, conceptual role determines the nature of the function between the world and a given mental symbol, even if everything we believe about the entities that fall under a given natural kind concept are up for revision.

Of course the challenge for any theory of narrow content is specifying, at least in principle, how we can separate which aspects of conceptual role are content determining and which are merely part of what we believe to be true about the entities a given mental symbol picks out. Otherwise we must endorse a holistic approach to narrow content. Moreover, if we can specify how to separate conceptual role into these two components, we bolster our confidence in a dual factor theory.

Fully grasping the implications of conceptual discontinuities, plus the nature of the bootstrapping processes that achieve them, provides one route into motivating narrow content and specifying which aspects of conceptual role are content determining. An appreciation of Quinian bootstrapping allows us to see how new primitive symbols are coined in the first place. At the outset of a bootstrapping episode, the concepts in placeholder structures have only narrow content; it is entirely given by concept role within the placeholder structure. I propose that those aspects of conceptual role in descendent structures that maintain that initial conceptual role, or derive, through conceptual change, from it, are among those aspects of conceptual role that determine narrow content. This proposal is consistent with others in the literature—that the most causally or inferentially central aspects of conceptual role determine narrow content. The constraint satisfaction modeling processes that enrich the content of placeholder structures maximize causal/inferential coherence. In sum, I appeal to to Quinian bootstrapping in my thinking about narrow content, as well as in my reply to Fodor’s radical concept nativism.

## 10. Concluding Remark

Explaining the human capacity for conceptual thought has animated philosophical debate since the time of the ancient Greeks and psychological debate since the dawn of experimental psychology in the 19$^{th}$ century. The many case studies in TOOC illustrate the interdependence of the projects of explaining the origin of concepts and understanding what concepts are.