Journal of Experimental & Theoretical Artificial Intelligence,
Vol. 17, No. 4, December 2005, 343–369

Taylor & Francis
Taylor & Francis Group

How we learn about things we don't already understand

DAVID LANDY* and ROBERT L. GOLDSTONE

Indiana University, Bloomington, IN, 47405, USA

(*Received February 2005; accepted May 2005*)

The computation-as-cognition metaphor requires that all cognitive objects are constructed from a fixed set of basic primitives; prominent models of cognition and perception try to provide that fixed set. Despite this effort, however, there are no extant computational models that can actually generate complex concepts and processes from simple and generic basic sets, and there are good reasons to wonder whether such models may be forthcoming. We suggest that one can have the benefits of computationalism without a commitment to fixed feature sets, by postulating processes that slowly develop special-purpose feature languages, from which knowledge is constructed. This provides an alternative to the fixed-model conception without radical anti-representationlism. Substantial evidence suggests that such feature development adaptation actually occurs in the perceptual learning that accompanies category learning. Given the existence of robust methods for novel feature creation, the assumption of a *fixed basis set* of primitives as psychologically necessary is at best premature. Methods of primitive construction include: (a) perceptual sensitization to physical stimuli; (b) unitization and differentiation of existing (non-psychological) stimulus elements into novel psychological primitives, guided by the current set of features; and (c) the intelligent selection of novel inputs, which in turn guides the automatic construction of new primitive concepts. Modelling the grounding of concepts as sensitivity to physical properties reframes the question of concept construction from the generation of an appropriate composition of sensations, to the tuning of detectors to appropriate circumstances.

Keywords: Novelty, Fixed-feature representations, Scientific discovery, Perceptual learning, Differentiation, Unitization, Cognition, Distributed cognition

*Corresponding author. Email: dlandy@cs.indiana.edu

Journal of Experimental & Theoretical Artificial Intelligence
ISSN 0952–813X print/ISSN 1362–3079 online
http://www.tandf.co.uk/journals
DOI: 10.1080/09528130500283832

1. Introduction

Nearly every model of intelligent behaviour traffics in concepts on some level. Successful models of abstract conceptual reasoning generally assume that concept identification is accomplished prior to abstract reasoning, and thus treats concept as unanalysed primitive tokens that can be reliably re-identified. On the other side, work on perception often ignores the high-level purposes which the act of perceiving is supposed to support (Marr 1982, Ullman 1984, Pylyshyn 1999).

These approaches to cognition gloss a difficult problem people face: since we explore novel domains for which we are evolutionarily unprepared, we must develop an understanding of the properties in terms of which a domain of knowledge can successfully be described, and we must of course do it *before* we acquire a rich understanding of the domain in question. In extreme cases like scientific exploration, domains of knowledge are constructed in which the relevant properties are not known by anyone prior to the construction process. In this case, the construction of an appropriate vocabulary of relevant concepts is a vital part of the cognitive task.

Consider the prominent computational model of analogical reasoning called structure mapping theory, as implemented in the structure mapping engine (SME) (Gentner 1983, Falkenhainer *et al*. 1990). In this model, analogies are typically drawn between situations involving very abstract high-level relations like 'revolves around'. The detection of this (surely phenomenologically complex) property is not dealt with, but is instead assumed to be handled by a prior system (which may, however, be directly instructed to search for the relation in a particular episode if an analogy reveals that such a relationship is likely). Perhaps more important, the decision to include this particular relation in the set of relevant properties is made beforehand, either by a designer or by an independent system. In practical systems based on SME which need to perform concept detection (e.g. to support a user interface), the task of identifying the concepts expressed in physical depictions is often deliberately simplified by requiring that the user respect a particular language of interaction (Forbus *et al*. 2003, 2004). The model PHINEAS (Falkenhainer 1988) uses structure mapping to explore a novel physical domain (heat flow) via analogies to known situations. To do so, PHINEAS depends on a previously available language of description well suited to qualitative descriptions of heat flow situations, but not obviously available ahead of time for a real cognitive agent. Furthermore, not only the description terms but also their generic similarity to analogous terms in water flow cases, are defined ahead of time for the system (This is hardly a novel comment: Falkenhainer specifically mentions the use of specialized feature languages as a prominent unexplained aspect of the model, as do (Chalmers *et al*. 1992).) More recently, research on the ambitious Digital Aristotle project (Friedland *et al*. 2004) relies on special-purpose languages designed specifically toward coverage of particular scientific domains. Digital Aristotle is intended to be a domain-generic scientific tool that can suggest connections and references between a user's research and research from other domains or areas. This work is still in its infancy, but in its opening stages, the Digital Aristotle relies on human-constructed special-purpose ontologies of concepts relevant to domains under its

coverage. While three different implementations all rely on general concepts, each uses topically specific basic concepts in its ontology as well.

In fact, practically oriented artificial intelligence models almost always employ special-purpose ontologies of primitive properties. Complex rules and entities are defined in terms of these properties. Since we can presume that natural problem-solvers are not pre-equipped with special-purpose languages for, say, chemistry labs (though see Fodor 1975, 1992), then to the extent that these models accurately reflect the concepts that human specialists use to understand their domains, we must wonder where these special-purpose languages come from. While one might try to take cognitive models to task for simply assuming feature languages which are specially geared toward the task being modelled and thus simplifying the real task facing a natural agent (Chalmers *et al.* 1992), we consider the use of specialized feature languages to be a very revealing instance of an intelligent design decision. The reason that models use specialized feature sets is, after all, that they make the task of learning much easier for the artificial agent. We suspect that in this case, art follows nature; real intelligent systems 'cheat' in exactly the same way that artificial systems do, by developing small languages of primitive features geared toward solving particular types of problems. Throughout the course of this paper, we will use the term 'language' to refer to a collection of elements and operations over those elements which act functionally as a special-purpose unit for some specific collection of purposes. Whether any particular primitive is part of a language may be a matter of degree, and languages need not be disjoint, but it is part of our story that many aspects of abstract human cognition can be divided roughly into different special-purpose languages, which behave more or less as a package. These languages are not necessarily public or spoken, but reflect (at an appropriate level of abstraction) the underlying cognitive processes. In particular, we are not implying an association between these languages and natural language, but rather characterizing these languages as something like formal computational languages. The question that we are concerned with is whether it is reasonable to presume that people have an underlying fixed-primitive language out of which all the other languages are built.

Anyone who has played with LEGO^{TM} understands the power of small, generic building blocks with specific combinatorial rules for constructing large representative structures. The combinatorial model has had a major impact on cognitive science and artificial intelligence over the last 40 years. The functional value, and also the cost, of employing varying basic languages is readily apparent. Generic languages that can express many different concepts and distinctions contain many distinctions that are irrelevant to particular tasks. Basic sets of primitives that are specially geared toward particular topics with more limited computational expressibility can more readily construct the most useful discriminations. PHINEAS, for instance, gains computational purchase on difficult problems by excluding much of the rich detail of real novel scientific situations from consideration—without this sort of simplification it could not practically function. In this manner, the (more or less) exclusive use of a small, custom feature set can greatly reduce the computational difficulty of learning a category or rule, as long as that rule can be well-discriminated in terms of those features. The source of special-purpose languages is heavily constrained by the assumption that cognitive operations are universally symbolic.

Computational cognitive operations always involve the combination of a primitive feature set into larger units (e.g. inference, or chunking methods of rule construction), so if the development of special purpose-languages is a computational operation, it must function hierarchically over some exceedingly general-purpose language. The dominant view thus requires that special-purpose languages are complexes of simpler and more general psychological languages; these 'lower-level' properties may be implemented in terms of still simpler features, in a cascading hierarchy of specificity, grounded in a base-level language that employs a fixed collection of psychological primitives. Several extant research programs aim to specify a fixed set of features which can be combined to generate all representations of visual objects (Biederman 1987), and semantic descriptions of scenes and sentences (Schank 1972, 1976, Schank and Kass 1988). Even if it is not specifically stated, however, the very assumption that cognition is computation—that all interesting cognitive activities involve the application of rules in a symbol system (Newell and Simon 1976)—entails a fixed set of primitives, or at least demands that any alterations to the primitive set are not cognitively interesting acts.

Despite its many merits, we believe that the computational perspective is importantly incomplete. Complex computational reasoning need not be implemented on top of a fixed set of generic primitive features. Instead, many cases of real abstract concept learning are best viewed as involving the construction of new psychological primitives. Concluding that cognition often combines primitive elements into compound units is not the same as concluding that all properties, or even all properties within a single modality or domain, are well described by a fixed set of primitive features. One can have the many clear benefits of computationalism without needing fixed feature sets, by postulating non-computational mechanisms that slowly alter feature languages and, in particular, are sensitive to systems of high-level categories, and adapt to discriminate them. The model of slowly-adapting primitive feature languages provides an alternative to the fixed-model conception without requiring radical anti-representationalism. Learning, we think, often involves the construction of psychological units that were unavailable before learning, and which transform the expressive capacities of the agent's psychological language. Cognitive processes like chunking alter what is tractably expressible, but we postulate processes that alter and extend what is expressible in principle in the cognitive language (i.e. using the basic terms and combinatorial rules) of the agent. Understanding the role of feature-construction in the learning of perceptual categories in controlled experiments can provide insight into the possible cognitive role of feature-construction in more abstract cognitive domains such as scientific and mathematical reasoning.

The task-dependent construction of expressive primitives is incompatible with the popular view that modal perception is prior to the demands of individual tasks, and does not depend on them. Models of low-level perception that produce scene descriptions in terms of putatively domain-general building generally assume either (a) that perception is independent of cognitive needs (pre-cognitive) and constrains what is learnable or (b) that perception assumes a well-understood prior task domain, so that the right divisions can be made in low-level perception. If we are right, then such stories are not so much wrong in their proposed building blocks, as simply missing a major part of the interesting story, e.g., how building blocks like this are constructed in the first place.

1.1. *What are variable-feature languages?*

To understand what we mean by feature creation, it is helpful to first analyse what we mean by 'feature'. By 'feature', we mean a psychological unit of perception or thought. 'Dimensions' are similarly psychological entities, but refer to a set of values that can be ordinally positioned. Brightness, then, is a psychological dimension only because it is processed as a unit. If luminance energy were not psychologically isolated then there would not be a (psychological) dimension of brightness reflecting this physical quantity. While a strong distinction is often made between features and dimensions, here we use the word 'feature' to refer to both.

Fixed-primitive methods of learning include chunking or combining primitives (Newell 1990), inference, induction-based learning, visual imagery (as accomplished, for example, Croft and Thagard 2001 and Davies 2004) and analogy (Mitchell 1993, Hummel and Holyoak 1996, 1997, Kokinov and Petrov 2001). These methods all work by compositionally combining symbolic elements. Much of the research in associative learning and neural-networks is also reasonably regarded as fixed-primitive, because much of that literature uses inputs and outputs which are already functionally (and task-specifically) encoded. The network's internal task is not then to construct symbols that correspond to world properties, but to find appropriate integrations of smaller languages. What these systems share is a fixed maximal expressive capacity; no entities can be created which cannot be expressed in principle from the underlying language. The basic set of primitives need not be small: Jerry Fodor notoriously argued for a very large set of primitives, so that roughly each word has a individual concept (though see Fodor 1998). Fixed primitive models, along with computationalism, pervade the cognitive study of concepts.

The dominance of the fixed-feature approach is surely bolstered by the paucity of alternatives. If distinctively cognitive activity comprises the application and construction of rules, productions, and new integrations of current psychological features into new complex terms, then the expressive capacity of a cognitive agent is fixed regardless of the techniques used to make new elements: none of these techniques can generate new primitive terms that were previously inexpressible in the language. From this perspective, then, the question is not whether there is a fixed primitive set, but which concepts are in it. Alternative viewpoints which lack this property, such as dynamic systems theory, largely entail giving up computationalism altogether (van Gelder 1995, Smith and Thelen 1996).

There is an answer to this 'what else' challenge: our alternative to fixed-primitive languages involves not giving up computationalism, but enriching it with mechanisms which allow the construction of new psychological primitives that are not just combinations of other known categories. There is significant evidence in the learning of perceptually defined categories that indicates that novel primitive psychological features are constructed, thereby increasing and altering the complex space of psychologically available parsings of certain visual scenes. We think that mechanisms similar to those important in perceptual learning could also be very useful in learning more abstract 'high-level' domains. Harnad has named the rooting of symbolic primitives in physical world-properties 'symbol grounding' (Harnad 1990, 2003); his approach differs from ours primarily in that while he emphasizes a role for symbol grounding in differentiating computational simulations from genuine intelligence, we are interested in actual cognitive effects on the processes and capacities

of symbolic systems. Our claim (which is compatible with Harnad's) is that the ways that symbols are grounded in the physical world have an important impact on the symbolic capacities, through non-computational mechanisms that construct feature languages and processes which are finely tuned to particular tasks.

The concept of variable-feature languages can be clarified by another example from the world of LEGO (though one may reasonably object that this is only a toy domain!): LEGO bricks come in two very distinct varieties. The first is the generic shapes that are good for many different models. These are typically called bricks, blocks or elements. LEGO sets in the 1970s and early 1980s contained mainly these generic, powerfully expressive pieces. However, as the LEGO domain branched out to novel domains like pirates and sailboats, and to specific fictional settings like Star Wars, pieces were produced that are both detailed and particular. These so-called 'cheat pieces' are very useful for representing specific objects or domains, but have far less general utility. For instance, sails cannot be satisfactorily constructed from generic LEGO pieces, but single-piece sails cannot be easily used for other purposes. Most models of primitive feature-sets are similarly geared toward particular tasks: and like the LEGO sets, they tend to change significantly when applied to new domains. Actually, we think this is a pretty good model for how primitive features work for people: for many tasks generic units work very well, but the detection of some categories requires (or greatly benefits from) the construction of novel features that were not previously part of the perceptual repertoire. This implies that there two different kinds of categorical learning: one in which features are combined in novel ways to correctly categorize an input set, and another in which novel features are constructed in a opportunistic way, to match the diagnostic properties of a category. We think that people are a bit less like children playing with LEGO, and more like the LEGO company—building new sets of LEGO to fill particular task-specific niches—than the usual rendition of the cognitive paradigm suggests.

2. Feature construction in perceptually grounded categories

We would like to be able to write a paragraph that runs something like this: the claim that novel perceptual features can be learned sounds murky, or even mystical, without the clarification that the novel features are always drawn from a larger, more expressive, more primitive language embodying the physical and pre-conceptual constraints on what can be incorporated into features in the first place. Features are not created out of nothing. They are built out of stimulus elements. The set of stimuli together with the properties of the specific receptors constrain the psychological features that are actually constructed: that is, they control the set of psychologically and physically possible features. Within that large set is a smaller collection of psychological primitives which a person has actually constructed. For most categorizations, it suffices to build complexes of available feature elements so as to match the diagnostic characteristics of some novel category. The construction of novel features is the adjustment of the space of available descriptions to match a novel domain with novel diagnostic features, which is a (heavily constrained) search through the larger space. We would like to write this, because rooting feature construction in a language that respected the innate perceptual constraints

of an agent would clarify and naturalize the construction of novel psychological primitives; indeed, much perceptual learning can be seen in this light. However, the adjustment of a large space of psychophysically possible features into a discrete, limited subset is just one of several likely mechanisms for novel feature construction. We will discuss some of these other methods in the next section, but for now, let us consider the evidence for feature construction that can, at least in principle, be taken to be a selection of a small set of important features from a very large fixed collection of basic primitives.

Features and dimensions are the units of perception and thought from which concepts and rules are constructed; we can now ask what physical aspects are bundled together into these psychological units. Features can be interpreted as packages of stimulus elements that are separated from other sets of elements and reflect the subjective organization of the whole stimulus into components.

Features can be revealed using several experimental operationalizations. If two pieces of physical information, X and Y, are packaged together in the same psychological feature and Z is not, then several empirical predictions follow. We predict that searching for X and Y simultaneously should be easier than simultaneously searching for X and Z (Treisman and Gelade 1980). We predict that searching for X should be affected by contextual variation to Y more than Z (Gauthier and Tarr 2002). We predict that categorization based on X should be slowed more by irrelevant variation to Y than Z (Garner 1974, 1976). It should be easier for people to attend to X and Y simultaneously than X and Z. All of these operationalizations tie in to the notion that X and Y are being processed together.

It is also noteworthy that all of these operationalizations imply a continuum of featurehood. There will be various degrees to which stimulus aspect Y intrudes upon or facilitates processing of X. Although we conceive of features as packages of stimulus components, we are not proposing that packages are completely discrete or mutually exclusive. Rather, they are packages in the same way that cliques can be circled in social networks or regions can be identified in brain neural networks. In all three domains, a unit (feature, clique or region) is characterized by relatively dense within-unit connectivity among elements and relatively sparse connectivity between elements within the unit and external elements. Features are useful idealizations because they capture the notion of elements that are densely interconnected, but it is important to recognize that (1) features (e.g. densely interconnected clusters) may exist at multiple levels of resolution, (2) elements processed as one feature may not have uniform interconnectivity to other elements of the same feature and (3) the internal integrity of different features may vary.

2.1. *Characterizing featural change*

Having characterized psychological features, we can now turn to the meaning of feature creation. By this account, feature creation simply involves alterations to the organization of stimulus elements into features. Figure 1 shows two ways that this can happen.

By unitization, stimulus elements (circles) that were originally processed into three features (ovals) come, with practice, to be processed by only two features. Elements that were originally processed separately are processed together (Shiffrin and Lightfoot 1997, Goldstone 2000). By differentiation, the same three-element object

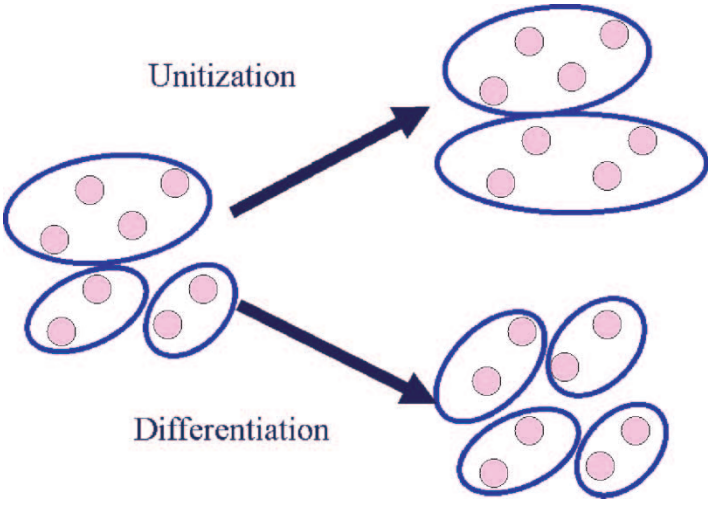

Figure 1. An abstract schema for differentiation and unitization.

comes to be processed into four features. Elements that were originally psychologically fused together become isolated (Smith and Kemler 1978, Smith *et al.* 1997, Goldstone and Steyvers 2001).

From the figure, it may appear that there are two separate, perhaps contradictory tracks for featural change. In fact, not only are unitization and differentiation compatible with each other, but they often occur simultaneously. They are compatible because both processes create appropriate sized units for a task. When elements co-vary together and their co-occurrence predicts an important categorization, the elements will tend to be unitized. If elements vary independently of one another and they are differentially relevant for categorizations, then the elements will tend to be differentiated. Accordingly, we do not support theories that propose monolithic developmental trends toward either increasingly unitized (Gauthier and Tarr 2002) or differentiated (Kemler and Smith 1978) representations. We believe that both occur, and furthermore, that the same learning algorithm can do both simultaneously (Goldstone 2003).

We hope that our characterization of featural change as reorganization of elements into processing units demystifies the process. Features are not created 'out of nothing.' They are built out of stimulus elements. A critic might respond, 'Then why is your account any different from the standard fixed-features approach in which primitive elements are combined in new arrangements to create object representations?' See (Schyns *et al.* 1998) for a full response to this potential objection. Here we will mention only two.

First, by our account, features are not (always) created from a set of psychological primitives. They are created from stimulus elements that often originally have no parsing in terms of psychological primitives. For example, people can create a 'saturation' detector that is relatively uninfluenced by brightness even if there was originally no detector that had this response profile (Burns and Shepp 1988). To be sure, if brightness and saturation affected a brain identically, then there would be no way to develop a detector that responded to only one of these properties.

However, as long as there are some differential effects of two properties, then increasingly differentiated detectors can emerge if the training encourages their isolation. The critic might counter 'But dimensions that are fused together at some point in perceptual processing can never be split later'. By analogy, once red ink has been poured into blue ink, there is no simple procedure for later isolating the blue ink. Fortunately, this analogy is misleading, and there are several computational models that can differentiate fused dimensions (Smith *et al*. 1997, Edelman 1999, Goldstone 2003). For example, competitive learning networks differentiate inputs into categories by developing specialized detectors for classes of stimuli (Rumelhart and Zipser 1985). Random detectors that are slightly more similar to an input than other detectors will learn to adapt themselves toward the input and will inhibit other detectors from doing so. The end result is that originally homogeneous detectors become differentiated and heterogeneous over training.

In addition, there are clear-cut cases where something like new perceptual devices are created. By becoming physically modified, systems can learn to represent properties that they were originally unable to represent. In evolutionary time, organisms developed ears sensitive to acoustic properties that no early organisms (e.g. bacteria) could detect. This is also possible within a system's own lifetime. The cybernetician Gordon Pask built a device that could create its own primitive feature detectors. It consisted of an array of electrodes partially immersed in an aqueous solution of metallic salts. Passing current through the electrodes grew dendritic metallic threads. Eventually the threads created bridges between the electrodes, which subsequently changed the behavioural repertoire of the device. Cariani (1993) reports that within a half a day, the system could be grown to be sensitive to a sound or magnetic field. With more time, the device could discriminate between two musical pitches. Similarly, there is good neurophysiological evidence that training can produce changes to early somatosensory, visual and auditory cortex (see Goldstone (1998) for a review). While these changes are not as radical as sprouting a new ear, they are existence proofs for how early perceptual devices can be systematically and physically altered by the environment to change their representational capacities.

2.2. *Unitization*

One result of category learning is the creation of perceptual units that combine stimulus components that are useful for the categorization. Such a process is one variety of the more general phenomenon of unitization, by which single functional units are constructed that are triggered when a complex configuration arises. Cattell (1886) invoked the notion of perceptual unitization to account for the advantage that he found for tachistoscopically presented words relative to non-words. Gestalt psychologists proposed the perceptual principle that objects will tend to be perceived in terms of components that have acquired familiarity (Koffka 1935). Weisstein and Harris (1974) found that briefly flashed line segments are more accurately identified when they are part of a set of lines forming a unitary object rather than an incoherent pattern. They interpreted this effect as showing that arrangements of lines can form configural patterns that are perceived before the individual lines are perceived. More recently Gauthier (1998) and Gauthier and Tarr (2002) found that prolonged experience with a novel object leads to a configural representation

of it that combines all of its parts into a single, viewpoint specific, functional unit. Their evidence for such a representation is that recognition of these familiarized objects improved considerably with practice, and was much more efficient when the object was in its customary upright form rather than inverted. Unitization has also been explored in the field of attention. Using a task where participants decided whether or not two visual objects were identical, Laberge (1973) found that when stimuli were unexpected, participants were faster at responding to actual letters than to letter-like controls. Furthermore, this difference diminished as the unfamiliar letter-like stimuli became more familiar over practice. He argued that the shape components of often-presented stimuli become processed as a single functional unit with practice. More recently, Czerwinski *et al*. (1992) have referred to a process of perceptual unitization in which conjunctions of stimulus features are bound together so that they become perceived as a single unit. Shiffrin and Lightfoot (1997) argued that separated line segments can become unitized following prolonged practice with the materials. Their evidence comes from the slopes relating the number of distracter elements to response time in a feature search task. When participants learned a conjunctive search task in which three line segments were needed to distinguish the target from distracters, impressive and prolonged decreases in search slopes were observed over 20 hour-long sessions. These prolonged decreases were not observed for a simple search task requiring attention to only one component.

Unitization is also important during the development of object perception. Newborn infants fail to integrate the separate regions of an object that is occluded (Slater *et al*. 1990). However, by 4.5 months of age, babies form the same interpretation of displays whether they are fully visible or occluded (Johnson 1997). This developed ability to integrate different parts into a single object representation depends on the featural similarity of these parts (Needham 1999).

In our own work on unitization, we (Goldstone 2000) gave participants extended practice learning the categorization shown in figure 2. In this categorization, a single object belongs to Category 1, and five very similar objects belong to Category 2. No single piece of the Category 1 doodle suffices to accurately categorize it because each piece is also present in several Category 2 doodles. Instead, all of its pieces must be considered. After 20 hours of practice with these stimuli we find that participants eventually can categorize the Category 1 doodle very accurately, and more quickly than would be predicted if they were explicitly combining separate pieces of information from the doodle together. Consistent with other work on perceptual unitization (Shiffrin and Lightfoot 1997, Gauthier 1998), we argue that one way of creating new perceptual building-blocks is to create something like a photograph-like mental image for highly familiar, complex configurations. Following this analogy, just as your local camera store does not charge more money for developing photographs of crowds than pictures of a single person, once a complex mental image has been formed, it does not require any more effort to process the unit than the components from which it was built.

2.3. *Dimension differentiation*

Selective attention is a critical component of adaptive learning, but it may not be the only process that dynamically alters the description of an object in a categorization task. A second candidate process is dimension differentiation, by which dimensions

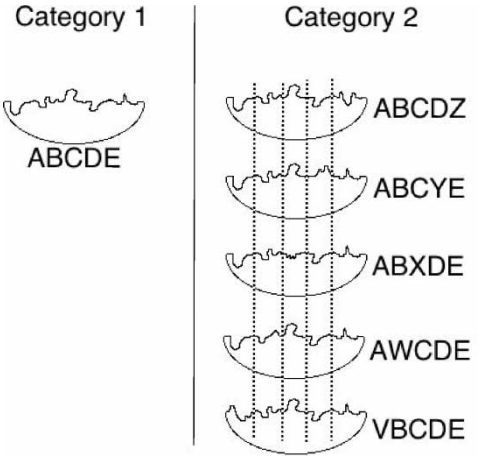

Figure 2. Stimuli used by Goldstone (2000). Each letter represents a particular stimulus segment, and each stimulus is composed of five segments. To categorize the item represented by 'ABCDE' as belonging to Category 1, it is necessary to process information associated with each of its segments because every single-segment distortion from ABCDE belongs in Category 2. The dashed lines in the right column were not part of the stimuli, but show the decomposition of the stimuli into their five components.

that are originally psychologically fused together become separated and isolated. Selective attention presumes that the different dimensions that make up a stimulus can be selectively attended. To increase attention to size but not colour, one must be able to isolate size differences from colour differences. In his classic research on stimulus integrality and separability, Garner argues that stimulus dimensions differ in how easily they can be isolated or extracted from each other (Garner 1974, 1976). Dimensions are said to be separable if it is possible to attend to one of the dimensions without attending to the other. Size and brightness are classic examples of separable dimensions; making a categorization on the basis of size is not significantly slowed if there is irrelevant variation on brightness. Dimensions are integral if variation along an irrelevant dimension cannot be ignored when trying to attend a relevant dimension. The classic examples of integral dimensions are saturation and brightness, where saturation is related to the amount of white mixed into a colour, and brightness is related to the amount of light coming off of a colour. For saturation and brightness, it is difficult to attend to only one of the dimensions (Burns and Shepp 1988, Melara *et al*. 1993).

From the above work distinguishing integral from separate dimensions, one might conclude that selective attention can proceed with separable but not integral dimensions. However, one interesting possibility is that category learning can, to some extent, change the status of dimensions, transforming dimensions that were originally integral into more separable dimensions. Experience may change the underlying representation of a pair of dimensions such that they come to be treated as relatively independent and non-interfering sources of variation that compose an object. Seeing that stimuli in a set vary along two orthogonal dimensions may allow the dimensions to be teased apart and isolated, particularly if the two dimensions are differentially diagnostic for categorization. There is developmental evidence that

dimensions that are easily isolated by adults, such as the brightness and size of a square, are treated as fused together for four-year old children (Kemler and Smith 1978). It is relatively difficult for children to decide whether two objects are identical on a particular dimension, but relatively easy for them to decide whether they are similar across many dimensions (Smith 1989). Children show considerable difficulty in tasks that require selective attention to one dimension while ignoring another, even if the dimensions are separable for adults (Smith and Kemler 1978). For example, children seem to be distracted by shape differences when they are instructed to make comparisons based on colour. Adjectives that refer to single dimensions are learned by children relatively slowly compared to nouns (Smith *et al*. 1997).

The developmental trend toward increasingly differentiated dimensions is echoed by adult training studies. Under certain circumstances, colour experts (art students and vision scientists) are better able to selectively attend to dimensions (e.g. hue, chroma, and value) that comprise colour than are non-experts (Burns and Shepp 1988). Goldstone (1994) has shown that people who learn a categorization in which saturation is relevant and brightness is irrelevant (or vice versa) can learn to perform the categorization accurately, and as a result of category learning, they develop a selectively heightened sensitivity at making saturation, relative to brightness, discriminations. That is, categorization training that makes one dimension diagnostic and another dimension non-diagnostic can serve to split apart these dimensions, even if they are traditionally considered to be integral dimensions. These training studies show that, to know how integral two dimensions are, one has to know something about the observer's history.

Goldstone and Steyvers (2001) have recently explored whether genuinely arbitrary dimensions can become isolated from each other. Their subjects first learned to group the 16 faces shown in figure 3 into categories that split the faces either horizontally or vertically into two groups with eight faces each. The faces varied along arbitrary dimensions that were created by morphing between randomly paired faces. Dimension A was formed by gradually blending from Face 1 to Face 2, while Dimension B was formed by gradually blending from Face 3 to Face 4. Each of the remaining faces is defined half by its value on Dimension A and half by its value on Dimension B. Results showed that: (a) people could easily learn either horizontal or vertical categorization rules; (b) once a categorization was learned, participants could effectively and automatically ignore variation along the irrelevant dimension; (c) only the category-relevant dimension became perceptually sensitized when participants were given a transfer same/different perceptual judgment task; and (d) there was positive transfer between categorization rules that presumed the same organization of faces into perceptual dimensions and negative transfer between rules that required cross-cutting, incompatible organizations. Together, these results strongly suggest that there is more to category learning than learning to selectively attend to existing dimensions. Perceptual learning also involves creating new dimensions that can then be selectively attended once created.

3. The role of feature construction in abstract reasoning

We have discussed the occasional but important occurrence of novel feature construction in the learning of perceptual categories; we next turn to high-level

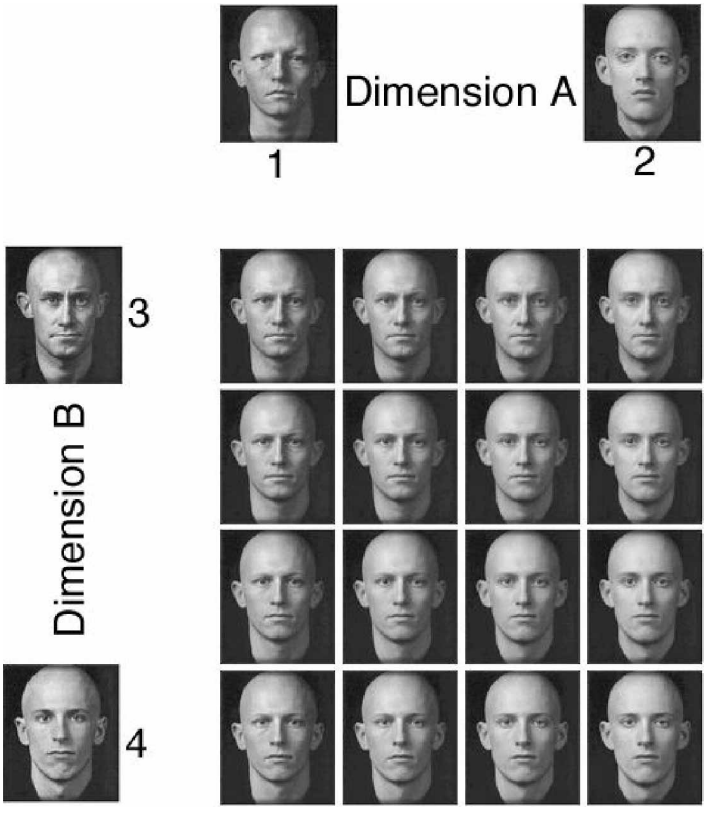

Figure 3. Stimuli used by Goldstone and Steyvers (2001). The four Faces 1, 2, 3 and 4 are blended in different proportions to create a 4 × 4 matrix of faces. The proportions of Faces 1 and 2 are negatively correlated such that the more of Face 1 present in one of the 16 centre faces, the less of Face 2 there will be. This negative correlation establishes Dimension A, and a similar negative correlation between Faces 3 and 4 establishes Dimension B. Each of the 16 centre faces is defined half by its value on Dimension A and half by its value on Dimension B.

conceptual tasks, and consider whether there is a similar role for primitive feature construction there. We will examine particularly the task of modelling scientific problem solving, because scientific understanding often involves the construction of novel contents, so it seems likely that novel conceptual primitives are often constructed in the course of routine scientific activity. Even in laboratory experiments with predefined perceptual categories, it is very difficult to establish that a particular instance of learning involves the construction of a novel primitive, and is not an instance of chunking prior primitives; it is similarly difficult to establish conclusively that any particular historical instance of scientific activity involves the construction of novel primitives. The goal of this section, therefore, is not to establish conclusively that variable feature languages are necessary to routine scientific activity, but just to argue that an alternative account to fixed feature sets exists, and to indicate where and how that alternative account might fit into an otherwise largely computationalist framework that depends on propositional compositions of well-chosen features. It is not sufficient for our case that new

primitive features be created—our case is that feature construction forms an interesting and vital part of the cognitive story of high-level categorization and reasoning. Though our case will only be indicative, not demonstrative, we find no reason to prefer models with small, fixed, generic building blocks to those that adapt their primitive features to fit novel problems.

The issue taken up here is distinct from the concern over incommensurability of scientific theories (Kuhn 1962, Carey 1988). There is general agreement that special-purpose high-level conceptual languages that represent two different theories may each be incapable of expressing each other, whether or not there is a more primitive psychological language that captures both. It is clear that special-purpose high-level conceptual languages play an important role in expert reasoning (Chase and Simon 1973, Chi and Glaser 1982, Smith *et al*. 1985, Carey 1988, Chi 1995, Nersessian 2002) and, furthermore, it is likely that language well-suited for expressing a particular theory or understanding of a domain will be unable to capture the distinctions important to very different practices and perspectives; the question here is over whether both languages are necessarily constructed out of a more basic language of generic psychologically available features. The computationalist assumption affords only two possibilities: either concepts are part of the primitive set of features available to an agent, or they are compositional constructions. Available models of concept creation such as model-based reasoning, imagistic reasoning, and analogy, all construct high-level concepts from fixed set of underlying features and relations. In contrast, we do not expect to find a generic set of psychological features that cleanly makes all necessary discriminations; instead, we think building the right feature discriminations is an ongoing, largely non-computational process, which is nevertheless a core piece of much scientific activity. This belief results from an acceptance, not a denial, of the assumption that scientific cognition is mechanistically continuous with everyday cognition. We think that building the right (custom) feature discriminations *is* an important part of everyday cognition. This perspective leads naturally to notions of what a cognitive model of particular historical scientific episodes should cover, and of the significance of aspects of scientific practice.

In the rest of this section, we will discuss specific ways that the conceptual tools made available by cognitive approaches to perceptual learning can inform models of scientific reasoning: in an imagined mathematical case, in Chi's model of radical conceptual change as ontological change, and finally in distributed-cognition models of scientific communities.

3.1. *Differentiation and unitization in conceptual reasoning*

There is little doubt that much of our arithmetic knowledge is interpreted in terms of the whole numbers and the basic operations of addition and multiplication. For high-school level mathematics, addition and numbers can be treated as unified entities, and laws can be considered, expressed, and no doubt represented in terms of them. That is, these elements form a natural feature set for modelling mathematical knowledge, and are used as the base language of knowledge expression in prominent theories of arithmetic learning, such as ACT-R (Lebiere 1998). However, when learning abstract algebra, the properties of the numbers and their operations are considerably redigested, and it seems quite likely that knowledge in this domain must

be structured in such a way that operations like addition over the naturals must be considered not as a primitive construct, but a defined and particular compound of more generic properties that are not readily apparent to the casual mathematical consumer, such as distributivity or associativity, or the presence or absence of identity of an element or inverse. When this reorganization occurs in the transition from an arithmetic perspective toward an algebraic one, we find that concepts that were unified entities in the initial language must be differentiated in the algebraic one. So for instance, an arithmetic understanding of addition does not distinguish between the operation of addition over the non-negative integers, and addition over the reals, but an understanding rooted in algebra treats these as separate concepts, since the latter is a group, but the former is a monoid. The initial arithmetic concept structure lacks the representational ability to express this difference, but because the structures are distinguished in the study of algebra, a new specialized language of features must be constructed. This situation mirrors the case of developed features in perceptual situations: the agent has a language which segments a conceptual situation poorly (for current purposes), but also is impacted differentially by different physical stimulus elements; our perceptual analogy suggests that in the abstract case initially homogeneous addition 'detectors' with only slight random variations in their response profiles might eventually become automatically tuned to different classes of addition. If so, this would form an important part of the cognitive story of learning algebra, but would not be the implementation of any computation on the part of the agent.

3.2. *Vestiges of the natural methods of feature creation*

Up to this point, our story runs something like this: non-computational perceptual mechanisms drive the construction of abstract primitive terms, which operate combinatorially in computational processes like analogy, chunking, generalization, inference, and so on. The learning mechanisms that govern the construction of high-level primitives share many of the properties of their more directly visual counterparts. Since we expect that primitive terms in languages operate much like feature detectors, the construction of novel primitives should often occur either through the differentiation of a term of the replaced theory into two, or the integration of two primitives into a single novel property. Several expectations follow: the development of expert conceptual schemes should not replace or eliminate prior naïve schemes, but will rather (at least sometimes) coexist alongside more traditional feature sets. As in perceptual categorization tasks, the high-level demands of the domain will play a significant role in guiding the construction of novel primitive features, but largely not through direct expression of rules or principles; instead, our expectation is that instruction regarding high-level conceptual primitives primarily aids learning by drawing attention to relevant diagnostic aspects of stimuli (because the segmentation is designed to construct perceptual units which correspond to diagnostic categories), and by controlling the character of training situations so as to optimize the performance of the processes of primitive construction. That is, instruction serves primarily to set up a categorical system which cannot be readily expressed from the terms available in the old language; this drives the adaptation of the underlying primitive set towards one more suited to discriminating the terms of the language the instructor is trying to teach.

From a computationalist perspective, supervision by instruction can help only if an agent can formulate a hypothesis in the first place. We suggest that the space of possible hypotheses can in fact be expanded opportunistically, if leveraged by instruction and perceptual-associative scaffolding. Part of our proposal is that segmenting situations appropriately is an important learned skill, and therefore that a significant part of the task of the science learner is to control the stimuli that they receive so as to optimize the automatic (non-rational) processes of their own perceptual system. Indeed, we suspect that an intuitive understanding of automatic perceptual processes forms part of the skill set of many successful scientists.

Tweney (1992) discusses the importance of Faraday's understanding of his own perceptual processes in Faraday's method for building an account of the effect of complex vibrations of a metal plate on metal filings resting on it. Tweney suggests that the internal representation of the visual patterns of the plate were not trained, and so for instance that the similarities between various such patterns could be readily detected by innate and domain-general processes. Mounting psychological evidence indicates, however, that perceptual learning and context both have a significant impact on similarity judgments of visual patterns (Goldstone *et al.* 1997). Faraday's experiments thus provide a valuable example of a scientific project in which results in perceptual learning are not only the basis of an appealing intuitive extension, but are in fact quite directly applicable.

Real perception of novel properties plays a significant role in some historical analyses of particular scientific traditions. Galison (1997), for example, emphasizes the role of training on judgments of similarities between photographic plate features in discoveries made by researchers working in the image tradition of high-energy particle physics. Galison notes the importance of visual similarity in persuading researchers in this very productive tradition. However, the visual 'similarity' under discussion is often not at all obvious without extensive practice and training (Pickering 1984); indeed, manuals of pictures of typical situations were produced to train the perceptual processes of aspiring scientists (Gentner *et al.* 1954). On our account, the deliberate exposure to stimuli that can, with an available and appropriate high-level categorization, automatically guide the construction of particular diagnostic combinations of physical sensitivities is a fundamental part of the cognitive task of this part of 20th-century particle physics. Unlike subjects of most categorization experiments, scientists expend a large effort organizing their own experiences (Cantor 1992). One significant result of this process is the construction via non-computational processes of the appropriate conceptual system; this construction is a significant cognitive activity, which occurs alongside computational combinations and law creation. If this analysis is correct, then cognitive studies of scientific activity can and should involve characterizing the processes that yield revisions of special-purpose conceptual languages.

So learning compositional special-purpose languages drives the construction of novel primitive features. The primitives generated by non-computational adjustments can, in turn, serve as generic hooks or terms from which general hypotheses, laws, and definitions can be constructed. As Harnad (1990) and Fodor (1998) argue, such primitives must be hooked up to the world, and this may support a notion of content more satisfactory than inferential role semantics. For modelling purposes, how the conceptual structure of an agent is hooked up to the world plays a role in the kinds of generalizations that structure can detect and support. That is, in order to

support the construction or evaluation of a general rule (even high-level), perceptual observations must be constructed that exhibit that rule. A set of primitives which is well-matched to a domain will permit the same computational resources to construct more successful constructions. So, in addition to providing symbolic tokens from which to construct language-like laws, hypotheses or properties, perceptual primitives provide the language of observation for an agent. Structural characteristics of psychological primitives may contribute to the evaluation of similarity between primitive terms, a process that is generally stipulative in models such as PHINEAS.

Non-computational perceptual processes have typically been studied with connectionist architectures. Because many connectionist architectures adaptively respond to imposed category structures, they are often used in modelling top-down influence on perception. Connectionist architectures are certainly useful and powerful tools with interesting applications to perceptual learning. However, the two are not interchangeable. First, connectionist architectures, like standard symbolic accounts, often assume that input has been heavily processed, and that relevant features have already been extracted before the stimulus ever reaches the 'input layer' (Elman 1990, 1993, Christiansen and Chater 1999). Thus, the task facing the network is one of combining semantically coded input statements. In models like Goldstone (2003), the input stimulus is taken to correspond to unprocessed features of the world, and the initial 'hidden layer' to correspond to the segmentation of the network. Here the network architecture and processes are constructed to closely reflect hypotheses about the nature of the constraints imposed by the agent's perceptual apparatus, instead of corresponding to a generic integration of pre-processed features. Thus, the initial segmentation layer adapts simultaneously to pressures coming from the world, the desired categorical system, and also the specific constraints imposed by the tools available in the act of perception. These aspects of the network are what make it an appropriate model of the interactions between perception and categorization.

3.3. *Radical ontological change as the result of perceptual learning*

Our discussion of the possible payoff of attention to detailed dynamics of perception in primitive construction has been abstract: at this point we will discuss in somewhat more detail one extant proposal within the cognitive studies of science that we feel might benefit particularly from a perceptual perspective: in particular, we consider Chi's proposal for the mechanisms governing radical conceptual change (Chi 1992, Chi and Hausmann 2003). Since our principle goal is not to provide a full account of this complex phenomenon, our coverage of the topic will be extremely brief, and we will not present a model of any underlying mechanisms. Our hope, rather, is that a brief consideration of Chi's proposal will help to illustrate the explanatory opportunities that a dual model of concepts, with a non-computational perceptual component enabling and influencing a computational component, can provide to extant purely computationalist process proposals.

Chi's ontological model runs like this: every concept is asserted to be connected to another, more generic concept (its ontological category); this set of categories forms a tree, whose root contains all concepts (Chi calls the root 'entity'). Since the structure of the tree is a psychological fact about an agent, how the tree is organized

depends on whose tree it is, but most people seem to share the deepest categories: thus, entities contain at a top level at least some concepts like MATERIAL, PROCESS and MENTAL STATE. And in turn, materials can be natural or artificial, processes can be intentional or physical, and mental states can be abstract or emotional. The relationship the ontological tree defines is containment, so that one path might contain the facts that 'anger is an emotional state', 'an emotional state is a mental state', and 'a mental state is a process'. Each ontological path from ENTITY constrains the types of attributes that a member of a concept may coherently embody: thus, material objects may have colour, but processes cannot. Chi calls the attributes that a concept's member might have 'ontological attributes'. Concepts inherit all the ontological attributes of their parents, and often have additional attributes of their own (which, in turn, will be passed to their descendents). However, it is important that these are only potential attributes, not attributes that are necessarily carried (or believed to be carried) by members of a particular category.

There are many facts about a concept that can change, on this sort of account. First, one might come to have a new belief about a concept's members, e.g. 'some squirrels are black'. Concepts can also change in such a way that the ontological tree is adjusted. First, you might have a new concept which integrates some common items: for instance, if someone learns the concept PINE TREE, then that concept will come between, say, the higher level concept TREE, and concepts SPRUCE or WHITE PINE. Another possibility is that a concept will leave its parent, and become the direct child of a grandparent or other ancestor; so one might at a certain point believe that ADDITION ON THE NATURALS is a child of ADDITION ON THE INTEGERS, which is a child of, perhaps ADDITION. After an introduction to abstract algebra, however, one might move ADDITION ON THE NATURALS to directly beneath the generic ADDITION. Another kind of ontological adjustment occurs when one concept moves from a particular branch to a separate branch; when, that is, the concept's parent becomes something that is not in the path from ENTITY to the concept. So if someone who thinks that WHALE is below FISH, which is below ANIMAL learns that whales are mammals, WHALE must be shifted to MAMMAL, which is not on the path from ENTITY to WHALE, nor is it a descendent of WHALE. In some cases like these, the ontological attributes of a concept may change, if it moves to a part of the tree with distinct ontological attributes. Chi argues that this kind of learning, called 'ontological change', is qualitatively different from other kinds of learning. In particular, ontological changes cannot be learned by induction, analogy, generalization, inference or any other psychological mechanism. The only way to make an ontological shift (more or less), is to be instructed to do so, and even in this case, such changes are slow, and often incomplete. Indeed, even physics experts seem to retain their naïve concept of force, in addition to gaining a new scientific concept of force as events. The psychological reality of ontological learning, as well as its distinctive characteristics, has extensive empirical support (Gelman 1988).

Within the context of this ontological model, then, Chi would like to equate the particularly difficult conceptual adjustments of physics and other domains with those that involve ontological change. And, in fact, the classes correspond quite well; for instance, conceiving of forces as a kind of process, rather than as a substance contained in a moving body, is very difficult for physics novices, and moving an

object from MATERIAL to PROCESS is a major ontological shift (since the lowest common ancestor is ENTITY). Let us take it for granted, then, that the qualitatively distinct facts mentioned above correspond quite well to those requiring difficult ontological changes, and consider why we might expect ontological shifts to be both difficult and qualitatively distinct from other kinds of learning. This is difficult to explain on a purely computationalist account; the tools of perceptual learning, however, offer a very natural account.

To review, the model of ontological change should account for the following characteristics, as naturally as possible: (a) ontological changes should be resistant to analogy, inference, generalization, and the attribution of properties; (b) ontological changes should be very difficult to learn under all circumstances, but less so under the influence of direct instruction; and (c) in many cases, ontological changes should be incomplete, in the sense that naïve intuitions should often be more or less unperturbed by the learning of new information.

The difficulty with a computationalist account is that it's hard to distinguish the ontological parent of a concept, from any of the other properties or rules which characterize a concept's role in an overall conceptual scheme. It is natural to want to represent the concept's parent as a primitive property of a concept; but then why shouldn't it be learned like every other property of a concept? We will review Chi's explanation of this aspect of things as it currently stands; we will then indicate why a familiarity with the tools of perceptual learning might prove useful in resolving this difficult issue.

Chi's resolution to the problem of making ontological change difficult on a computational account is to contend that it is, in fact, not difficult at all. That is, making an ontological adjustment is not difficult, if directly instructed to do so; what is difficult, on her account, is first, deciding that ontological recategorization is necessary, and second, dealing with the consequences of making such an adjustment. Ontological changes are, according to Chi, relatively rare in learning; therefore, it is not likely that a learner will try making an ontological adjustment. What is more, a learner may not have a thorough understanding of the new parent concept, which will make learning much more difficult: so for instance, learning that heat transfer is an emergent property of many small-scale events will be very difficult, unless emergent properties are already well understood. Even if the learner does understand the new concept, it may be very difficult to understand all of the implications of a new categorization, because each belief held about the old concept makes use of ontological attributes unavailable in the new position. Finally, it may often be the case that, rather than altering the parent of an existing concept, one constructs a brand new concept in the appropriate new place, and at some point attaches the known word to the new concept in place of the old. Chi observes that FORCE, for instance, may really refer to two separate concepts, one of an impetus (a type of MATERIAL) and another separate expert concept (a child of PROCESS).

The computational account given above does a good job of accounting for the phenomenon, but involves a certain amount of awkwardness and implausibility. First of all, it is not at all clear that ontological recategorizations are much less common than other kinds of learning. Chi provides no evidence for this claim, appealing to our intuition that ontological changes are rare. However, the opposite seems to be true; learning novel ontologies is extremely common, and very often involves lateral movements and the addition or deletion of ontogical properties.

Most people must learn, for instance, that charcoal and diamonds are both varieties of carbon; that genes are a kind of protein (and proteins are not a kind of vitamin); that whales and dolphins are fish, and tadpoles are not. Scientific taxonomies of natural kinds differ significantly from most naïve notions, and are typically taught in school; mushrooms are a kind of fungus and not a kind of plant, for instance, and sponges are not plants but animals. Learning technological systems requires many ontological changes: a Corolla is a kind of Toyota which is a kind of Japanese car which is a kind of car which is just a kind of vehicle. Nylon is a kind of plastic that is a kind of petroleum, which is, of all things, an organic material. Some rubber comes from a rubber plant, and so is, again, a kind of plant. Legal systems also demand lateral ontological change: ketchup is a vegetable, although tomatoes are a kind of fruit (learning that fruits are vegetables also requires an lateral ontological shift, and a shift in ontological properties); DVD burning may or may not be a kind of fair use; the right to abortion is currently a variety of the parent concept, CONSTITUTIONAL RIGHT. Many of these shifts involve small ontological attribute changes. For example, it makes sense to ask if the right to abortion is a right in the United States, only if it is represented as a sub-concept of CONSTITUTIONAL RIGHT. For someone who represents abortion rights as a type of basic human right, it makes no sense to attribute a nationality. In short, it seems that lateral ontological shifts happen frequently enough to call into question a claim that ontological change is ignored as a deliberate strategy. To be sure, none of these ontological changes are as fundamental as the change from MATERIAL to PROCESS; but rarity of lateral ontological change seems to be false (or at least undersupported) on the face of it, and it is rarity of ontological change that was supposed to account for the qualitative differences in the kinds of instruction that can generate ontological changes.

The worry is pressing; if ontological change really is (just) another property of concepts, and is not so rare as to be implausible, then it is not clear why learning an attribute that is incompatible with the current ontological position does not induce, more or less immediately, an ontological change. That is, ontological attributes, since they are disjoint across lateral positions, can entail that a miscategorization has occurred; if they do not cause recategorizations, then it must be that ontological changes are somehow inhibited. But this inhibition was supposed to fall out of the rarity of the step. If ontological changes are neither particularly rare nor particularly hard, it is very odd to suppose that people ignore standard forms of evidence, and require direct instruction. Indeed, even if the alteration of ontological attributes is rare, since it is easily inferred from the application by a teacher of nonsensical attributes, it is odd that learners don't take advantage of this information.

Chi also contends that ontological change is difficult because it implies that most or all beliefs involving the relevant concept must be reinterpreted. For instance, when one learns that heat is not a fluid but an emergent process, then one will need to revisit any beliefs one previously had regarding its viscosity. Though this is true, it is also the case that non-ontological belief changes may, and sometimes do, require radical revisions of global beliefs; this is the famous frame problem. Once again, there do not seem to be any qualitative differences between ontological change and other kinds of learning.

Consider now an alternate proposal, informed by a notion of primitive feature construction. On this account, each category is characterized by two separate sets of properties: the first are attributes and properties that an agent attributes to members

of that class, the second are intrinsic properties of a perceptual system that detects the presence of a primitive term. For simplicity of exposition, we will imagine that primitive detection occurs through a basic neural network connecting stimulus properties to the spoken language the agent is learning. On our model, then, two genuinely separate types of learning really do occur: laws, rules, and beliefs are learned through traditional psychological–computational mechanisms like analogy and inference. These computational mechanisms operate over basic primitives, whose intrinsic properties are learned on the basis of statistical and associative evidence (guided by perceptual apparatus). Ontological structure, on this account, is a fact 'about' a symbol, not a fact which contains a symbol. Beliefs or properties are compound statements referring to symbols; on the scheme we're offering, ontological category is a fact about the intrinsic character of a symbol that guides that property's detection. On this account, ontological categories fall out of the similarity space of the internal structure of the networks that recognize them.

Clusters of similar categories that fit particular contexts or into particular roles fall naturally out of many connectionist tasks. Figure 4 shows one such cluster diagram produced by a network constructed for Landy (2004).

In artificial and natural language studies, such networks often naturally induce trees that distinguish the primary categories of the languages on which they are trained (e.g., nouns yield a reliably distinct kind of representation from verbs

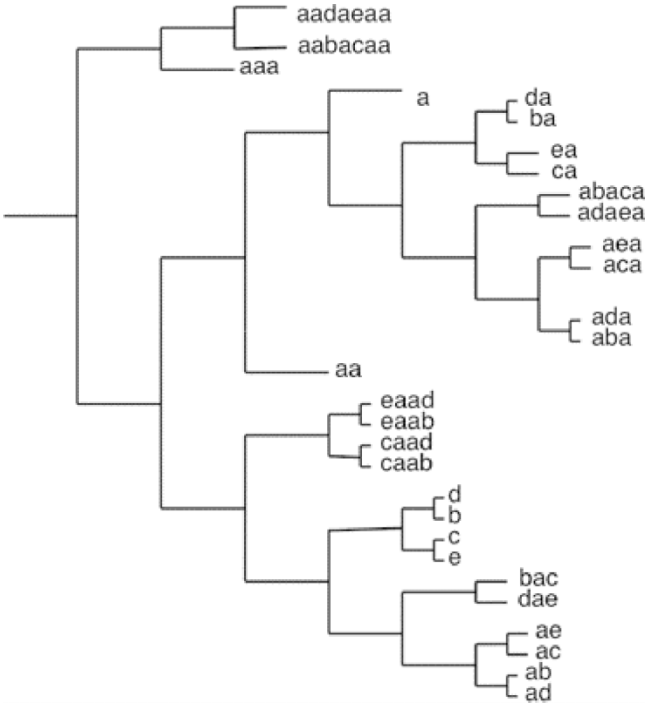

Figure 4. A cluster diagram from Landy (2004; this diagram appeared only in the talk). The language in this case was {a{bac|dae}a}*. Structurally similar elements are located close to one another.

(Elman 1995)). Similarly, ontological attributes are treated as statistical facts about the associative similarity of different primitives.

How does this account resolve the peculiarities in the computational approach to ontological change? First, if ontological learning is a kind of perceptual learning, then it is not at all surprising that naïve versions of concepts like force are typically retained. Differentiation is, as we have emphasized, a common perceptual operation. Second, the identification of ontological adjustment with the adaptation of the internal structure of a high-level primitive gives a natural explanation to the fact that property adjustment is responsive to different kinds of instruction than ontological category; they are different processes, working over very different kinds of representations, so it is not unreasonable that different methods of correction are effective for each case. Finally, this account explains why ontological changes are typically more amenable to direct instruction than to other forms of learning. The ontological category is not part of the externally represented properties of an entity. Rather, the phrase 'ontological category' describes a fact about the internal structure of the detector. This internal structural fact contributes to the combinatorial system by providing context-sensitive similarities between terms, and by informing and licensing particular conceptual combinations; but it is not overtly expressed in the language itself. As such, it is not directly correctable, in the way that a property, for example, may be. However, ontological category is taught by the process of training a subject to use a new categorical system. Thus, direct instruction that force is a process and not a substance will be effective because such instruction provides a new set of external categories that place force closer to other processes. However, this process will at best speed up an intrinsically extended process of adjustment, and is unlikely, even in best case, to make fundamental ontological conceptual changes less than difficult. However, ontological changes between nearby categories, such as FISH and MAMMAL, will be relatively easy, simply because the concepts are already represented similarly. On the clustering account, this similarity of representation is why they are nearby on the ontological tree in the first place.

Ontological change is an impressive account of radical conceptual changes in development. By taking a perceptual learning perspective, many perplexing features of the computational account of ontological learning become quite natural. We have not given a full account of ontological change or development, nor is that our purpose here. Instead, we indicate the value of non-computationalist perceptual learning in providing insight into the standard cognitive problems of scientific reasoning. Such learning clearly occurs, and an understanding of its characteristic properties can only enhance the capacity of cognitive studies of science to explain and discuss human behaviour.

3.4. *Perception in distributed cognitive agents*

The individual scientist is not the only cognitive agent on today's scientific scene. Considerable research suggests that distributed collections of scientists, institutions and tools or instruments may effectively constitute a single cognitive agent (Hutchins 1995, Clark and Chalmers 1998, Giere 2003, Nersessian *et al*. 2003). This method of analysis reveals further suggestive connections between feature construction in perception and the activities of a larger-scale cognitive agent. That experimental practices are indeed heavily guided by theoretical demands forms a prominent part

of modern analyses of historical situations (Pickering 1984, Kohler 1994). One general virtue of including the creation of high-level custom-purpose languages cashed out in practices and devices for the organization of physical stimuli in the cognitive story of science is the potential for increased compatibility with sophisticated and illuminating historical and sociological analyses of science, and for extending and deepening the analogy between experiment and theory in scientific communities and the processes of conceptual development in individuals.

If we consider extended social organizations such as laboratories as our fundamental cognitive unit, then the possibilities for non-computational construction of novel sensitivities become even more central. For instance, we observed that a saturation-only detector could not be constructed if saturation and brightness affected the cognizer identically. For unadorned individuals, this may be right: but research groups generally do construct novel sensitivities to aspects of their tasks.

We said in the previous section that we could not endorse the conservative-sounding claim that if two stimuli did not differentially impact an agent, then those stimuli could not be differentiated in principle. Here's why: cognitive agents engaged in processes of discovery, unlike subjects in perceptual learning studies, actively alter their own perceptual experiences. By the construction of measuring devices and exploratory tools and practices, scientific agents can circumvent limitations on the sensitivity of their measuring devices. The impact that the phases of the moons of Jupiter made on Galileo was greatly amplified by the use of a telescope. The construction of a novel instrument is often deliberately performed in order to differentiate between two possible physical situations that current techniques cannot distinguish; this is the same kind of process that governs differentiation in perceptual learning. In perceptual learning, the constraints of the detection system govern features which can be built, while in tool-building, technological and sociological considerations constrain what new devices, and therefore what new featural sensitivities may be constructed. However, the construction of new instruments does not just amplify existing differences, but may create entirely new sensitivities. For instance, the development of photographic plates incidentally produced a sensitivity to X-ray stimuli. In this sense, perception in high-level cognition can be more like the development of ears than the development of new feature detection units. These processes are an essential part of the cognitive story of scientific activity.

4. Discussion

Despite the large support for intrinsic primitive representational systems, systems that develop new psychological primitives in response to specific tasks have had considerable success in modelling many aspects of categorization and category learning. While they are only easily verified in laboratory experiments, it is likely that novel features are also constructed in the course of high-level conceptual learning, both in individual people, and also perhaps in distributed cognitive systems. Task or domain-specific primitive properties geared toward specific purposes are generated during the exploration of new domains through methods that are not usefully thought of as recombinations of an underlying constant vocabulary of primitive psychological elements. Using learning methods like competitive learning, which exacerbate small

differences in known categories through repeated exposure to diagnostic situations, and more complex and deliberate methods of tool and experiment construction, novel sensitivities can be constructed when needed.

Many approaches that deny the cognition-is-computation hypothesis further deny that computational models are a useful way to explore the mind (Searle 1980, van Gelder 1995); following Uhr (1978), Harnad (1990), Schyns *et al*. (1998), the approach to cognition discussed here endorses both the notion that cognition often is best viewed as the mental processing of symbols in a specialized formal language, and also that formal models of cognitive processes are nearly universal. Indeed, the models described throughout this paper have been formalized and implemented in computer programs. Essentially, what we deny is that specialized concepts are generally rooted in or composed of other, more basic concepts. Concepts are instead grounded in real perception and manipulation of a real world; this manipulation is surely guided by pre-existing concepts, and pre-existing concepts are often the basis for newly constructed ones, but not through processes which are best modelled by conceiving of the mind as a computer operating over abstract symbols. Dealing appropriately with the importance of concept formation and relation creation to scientific discovery and rule construction is a necessary step in modelling the cognitive aspects of this complex high-level activity.

The approach to the interaction between associative modelling and rule-based reasoning taken here is different than some which have been presented, in that rather than postulating two adjacent or competing systems, with separate components, we suggest two interacting, complementary systems If true, the construction of task-specific psychological primitives offers profound opportunities to researchers interested in modelling, either by a computer-based simulation, or in a psychological or sociological context, the cognitive aspects of scientific behaviours.

References

I. Biederman, ''Recognition-by-components: a theory of human image understanding'', *Psychological Review*, 94(2), pp. 115–147, 1987.

B. Burns and B.E. Shepp, ''Dimensional interactions and the structure of psychological space: The representation of hue, saturation, and brightness'', *Perception and Psychophysics*, 43, pp. 494–507, 1988.

S. Carey, ''Conceptual differences between children and adults'', *Mind and Language*, 3, pp. 167–181, 1988.

P. Cariani, ''To evolve an ear'', *Systems Research*, 10, pp. 19–33, 1993.

D. Chalmers, R.M. French and D. Hofstadter, ''High-level perception, representation and analogy: a critique of artificial intelligence methodology'', *Journal of Experimental and Theoretical Artificial Intelligence*, 4, pp. 185–211, 1992.

G. Cantor, ''How Michael Faraday brought law and order to the West End of London'', *Physis*, 29(1), pp. 187–203, 1992.

J.M. Cattell, ''The time it takes to see and name objects'', *Mind*, 11, pp. 63–65, 1886.

W.G. Chase and H.A. Simon, ''The mind's eye in chess'', in *Visual Information Processing*, W.G. Chase, Ed., New York, Academic Press, 1973.

M. Chi, ''Conceptual change within and across ontological categories: examples from learning and discovery in science'', in *Cognitive Models of Science*, R.N. Giere, Ed., Minneapolis, MN: University of Minnesota Press, 1992, pp. 129–186.

M. Chi, ''Experts' vs. Novices' Knowledge'', *Current Contents*, 15, 12, 1995.

M. Chi and R. Glaser, ''Categorization and representation of physics problems by experts and novices'', *Cognitive Science*, 5, pp. 121–152, 1982.

M. Chi and R. Hausmann, ''Do radical discoveries require ontological shifts?'', in *International Handbook on Innovation 3*, L.V. Shavinina and R. Sternberg, Eds, New York: Elsevier, 2003, pp. 430–444.

M. Christiansen and N. Chater, "Toward a connectionist model of recursion in human linguistic performance", *Cognitive Science*, 23(2), pp. 157–205, 1999.
A. Clark and D. Chalmers, "The extended mind", *Analysis*, 58, pp. 10–23, 1998.
D. Croft and P. Thagard, "Dynamic imagery: a computational model of motion and visual analogy", in *Model-based reasoning: Scientific discovery*, *technological innovation*, *values*, L. Magnani, Ed., New York: Kluwer/Plenum, 2002.
M. Czerwinski, N. Lightfoot and R.M., Shiffrin, "Automatization and training in visual search", *The American Journal of Psychology*, 105, pp. 271–315, 1992.
J. Davies, "Constructive adaptive visual analogy", PhD thesis Georgia Institute of Technology, 2004.
S. Edelman, *Representation and Recognition in Vision*, Cambridge, MA, The MIT Press, 1999.
J.L. Elman, "Finding structure in time", *Cognitive Science*, 14, pp. 179–211, 1990.
J.L. Elman, "Learning and development in neural networks: the importance of starting small", *Cognition*, 48(1), pp. 71–99, 1993.
J.L. Elman, *Language as a Dynamical System*, Cambridge, MA: The MIT Press, 1995.
B. Falkenhainer, "Learning from physical analogies: a study in analogy and the explanation process", PhD thesis, Urbana-Champaign, University of Illinois, 1988.
B. Falkenhainer, K.D. Forbus and D. Gentner, "The structure mapping engine: algorithm and examples", *Artificial Intelligence*, 41, pp. 1–63, 1990.
J. Fodor, *The Language of Thought*, New York, NY: Crowell, 1975.
J. Fodor, *A Theory of Content and Other Essays*, Cambridge, MA: MIT Press, 1992.
J. Fodor, *Concepts: Where Cognitive Science Went Wrong*, Oxford: Oxford University Press, 1998.
K. Forbus, K. Lockwood, M. Klenk, E. Tomai and J. Usher, "Open-domain sketch understanding: The nuSketch approach", in *Proceedings of the AAAI Fall Symposium on Making Pen-based Interaction Intelligent and Natural*, Arlington, VA, pp. 58–64, 2004.
K. Forbus, E. Tomai and J. Usher, "Qualitative spatial reasoning for visual grouping in sketches", in *Proceedings of the 17th International Workshop on Qualitative Reasoning*, Brasilia, Brazil, 2003.
N.S. Friedland, P.G. Allen, G. Matthews, M. Witbrock, D. Baxter, J. Curtis, J. Shepard, P. Miraglia, J. Angele, S. Staab, E. Moench, H. Oppermann, D. Wenke, D. Israel, V. Chaudhri, B. Porter, K. Barker, J. Fan, S.Y. Chaw, P. Yeh, D. Tecuci and P. Clark, "Project halo: Towards a digital Aristotle", *AI Magazine*, 25(4), 2004.
P. Galison, *Image and Logic*, Chicago: University of Chicago Press, 1997.
W.R. Garner, *The Processing of Information and Structure*, Potomac, MD: Lawrence Erlbaum, 1974.
W.R. Garner, "Interaction of stimulus dimensions in concept and choice processes", *Cognitive Psychology*, 8, pp. 98–123, 1976.
I. Gauthier and M.J. Tarr, "Becoming a 'Greeble' expert: exploring mechanisms for face recognition", *Vision Research*, 37, pp. 1673–1682, 1997.
I. Gauthier and M.J. Tarr, "Unraveling mechanisms for expert object recognition: bridging brain activity and behavior", *Journal of Experimental Psychology: Human Perception & Performance*, 28, pp. 431–446, 2002.
I. Gauthier, P. Williams, M.J. Tarr and J. Tanaka, "Training 'greeble' experts: a framework for studying expert object recognition processes", *Vision Research*, 38, pp. 2401–2428, 1998.
S.A. Gelman, "The development of induction within natural kind and artifact categories", *Cognitive Psychology*, 20(1), pp. 65–95, 1988.
D. Gentner, "Structure-mapping: a theoretical framework for analogy", *Cognitive Science*, 7, pp. 155–170, 1983.
W. Gentner and H. Maier-Leibnitz, *An Atlas of Typical Expansion Chamber Photographs*, New York, NY: Wiley, 1954.
R.N. Giere, "The role of computation in scientific cognition", *Journal of Experimental and Theoretical Artificial Intelligence*, 15, pp. 195–202, 2003.
R.L. Goldstone, "Perceptual learning", *Annual Review of Psychology*, 49, pp. 585–612.
R.L. Goldstone, "Unitization during category learning", *Journal of Experimental Psychology: Human Perception and Performance*, 26, pp. 86–112, 2000.
R.L. Goldstone, "Learning to perceive while perceiving to learn", in *Perceptual Organization in Vision: Behavioral and Neural Perspectives*, R. Kimchi, M. Behrmann and C. Olson, Eds, Mahwah, NJ: Lawrence Erlbaum Associates, 2003, pp. 233–278.
R.L. Goldstone, D.L. Medin and J. Halberstadt, "Similarity in context", *Memory and Cognition*, 25, pp. 237–255, 1997.
R.L. Goldstone and J.Y. Son, "The transfer of scientific principles using concrete and idealized simulations", *The Journal of the Learning Sciences*, 14(1), pp. 69–110, 2005.
R.L. Goldstone and M. Steyvers, "The sensitization and differentiation of dimensions during category learning", *Journal of Experimental Psychology: Human Perception & Performance: General*, 130, pp. 116–139, 2001.
S. Harnad, "The symbol grounding problem", *Physica D*, 42, pp. 335–346, 1990.

S. Harnad, *Cognition is Categorization*, UQaM Summer Institute in Cognitive Sciences on Categorization, 2003.
J. Hummel and K.J. Holyoak, "LISA: A Computational Model of Analogical Inference and Schema Induction", in *The Proceedings of the Eighteenth Annual Conference of the Cognitive Cognitive Science* Society, Hillsdale, NJ: Erlbaum, 1996, pp. 352–357.
J. Hummel and K.J. Holyoak, "LISA: A computational model of analogical inference and schema Induction", *Psychological Review*, 1997.
E. Hutchins, *Cognition in the Wild*, Cambridge, MA: MIT Press, 1995.
S.P. Johnson, "Young infants' perception of object unity: Implications for development of attentional and cognitive skills", *Current Directions in Psychological Science*, 6, pp. 5–11, 1997.
D.G. Kemler and L.B. Smith, "Is there a developmental trend from integrality to separability in perception?", *Journal of Experimental Child Psychology*, 26, pp. 498–507, 1978.
K. Koffka, *Principles of Gestalt Psychology*, New York: Harcourt Brace, 1935.
R.E. Kohler, *Lords of the Fly: Drosophila Genetics and the Experimental Life*, Chicago: University of Chicago Press, 1994.
B. Kokinov and A.A. Petrov, "Integration of Memory and Reasoning in Analogy-Making: The AMBR Model", in *Analogy: Perspectives from Cognitive Science*, D. Gentner, K. Holyoak and B. Kokinov, Eds, Cambridge, MA: MIT Press, 2001.
T. Kuhn, *The Structure of Scientific Revolutions*, Chicago, University of Chicago Press, 1962.
D. LaBerge, "Attention and the measurement of perceptual learning", *Memory and Cognition*, 1, pp. 268–276, 1973.
D. Landy, "Recurrent Representation Reinterpreted", in *Proceedings of The AAAI Fall Symposium on Connectionism and Compositionality*, Arlington, VA, 2004, pp. 40–43.
C. Lebiere, "The dynamics of cognition: An ACT-R model of cognitive arithmetic", *Computer Science*, Pittsburgh: Carnegie Mellon University, 1998.
D. Marr, *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*, San Francisco: W. H. Freeman and Company, 1982.
R.D. Melara, L.E. Marks and B.C. Potts, "Primacy of dimensions in color perception", *Journal of Experimental Psychology: Human perception and performance*, 19, pp. 1082–1104, 1993.
M. Mitchell, *Analogy-Making as Perception: A Computer Model*, Cambridge, MA: MIT Press, 1993.
A. Needham, "The role of shape in 4-month-old infants? Segregation of adjacent objects", *Infant Behavior and Development*, 22, pp. 161–178, 1999.
N. Nersessian, "Maxwell and 'the method of physical analogy': Model-based reasoning, generic abstraction, and conceptual change", in *Essays in the History and Philosophy of Science and Mathematics*, D. Malamêt, Ed., LaSalle, IL: Open Court, 2002, pp. 126–166.
N.J. Nersessian, E. Kurz-Milcke, W.C. Newstetter and J. Davies, "Research Laboratories as Evolving Distributed Cognitive Systems", in *Proceedings of The Twenty-fifth Annual Conference of the Cognitive Science Society*, A. Markman and L. Barsalou, Eds, Erlbaum, 2003, pp. 857–862.
A. Newell, *Unified Theories of Cognition*, Cambridge, MA: Harvard University Press, 1990.
A. Newell and H.A. Simon, "Computer science as empirical inquiry: symbols and search", *Communications of the ACM*, 19(3), pp. 113–126, 1976.
A. Pickering, *Constructing Quarks: A Sociological History of Particle Physics*, Chicago: University of Chicago Press, 1984.
Z. Pylyshyn, "Is vision continuous with cognition? The Case for Cognitive Impenetrability of Visual Perception", *Behavioral and Brain Sciences*, 22(3), pp. 341–423, 1999.
D.E. Rumelhart and D. Zipser, "Feature discovery by competitive learning", *Cognitive Science*, 9, pp. 75–112, 1985.
R. Schank and A. Kass, "Knowledge Representation in People and Machines", in *Meaning and Mental Representations*, U. Eco, M. Santambrogio and P. Violi, Eds, Bloomington, IN: Indiana University Press, 1988, pp. 181–200.
R.C. Schank, "Conceptual dependency: a theory of natural language", *Cognitive Psychology*, 3(4), pp. 532–631.
R.C. Schank, *Conceptual Information Processing*, New York: Elsevier Science, 1976.
P.G. Schyns, R.L. Goldstone and E. Thibault, "Development of features in object concepts", *Behavioral and Brain Sciences*, 21(1), pp. 1–54, 1998.
J. Searle, "Minds, brains, and programs", *Behavioral and Brain Sciences*, 3(3), pp. 417–457.
R.M. Shiffrin and N. Lightfoot, "Perceptual learning of alphanumeric-like characters", in *The Psychology of Learning and Motivation*, *Volume 36*, P.G. Schyns, R.L. Goldstone and D.L. Medin, Eds, San Diego: Academic Press, 1997, pp. 45–82.
A. Slater, V. Morison, M. Somers, A. Mattock, E. Brown and D. Taylor, "Newborn and older infants' perception of partly occluded objects", *Infant Behavior and Devleopment*, 13, pp. 33–49, 1990.
C. Smith, S. Carey and M. Wiser, "On differentiation: a case study of the development of size, weight, and density", *Cognition*, 21(3), pp. 177–237, 1985.

L.B. Smith, ''From global similarity to kinds of similarity: The construction of dimensions in development'', in *Similarity and Analogical Reasoning*, S. Vosniadou and A. Ortony, Eds, Cambridge: Cambridge University Press, 1989, pp. 146–178.
L.B. Smith and D.G. Kemler, ''Levels of experienced dimensionality in children and adults'', *Cognitive Psychology*, 10, pp. 502–532, 1978.
L.B. Smith and E. Thelen, Eds, *A Dyanamic Systems Approach to the Development of Cognition and Action*, Cambridge: MIT Press, 1996.
L.B. Smith, M. Gasser and C.M. Sandhofer, ''Learning to talk about the properties of objects: A network model of the development of dimensions'', in P.G. Schyns, R.L. Goldstone and D.L. Medin, Eds, *The Psychology of Learning and Motivation Volume 36*, San Diego: Academic Press, 1997, pp. 219–255.
A. Treisman and G. Gelade, ''A feature-integration theory of attention'', *Cognitive Psychology*, 12, pp. 97–136, 1980.
R.D. Tweney, ''Stopping time: Faraday and the scientific creation of perceptual order'', *Physis*, 29(1), pp. 149–164, 1992.
L. Uhr, ''Tryouts toward the production of thought'', in *Perception and Cognition Issues in the Foundations of Psychology*, C.W. Savage, Ed., Minneapolis, University of Minnesota Press, 1978, pp. 327–364.
S. Ullman, ''Visual routines'', *Cognition*, 18, pp. 97–159, 1984.
T. van Gelder, ''What might cognition be if not computation?'', *Journal of Philosophy*, 91, pp. 345–381, 1995.
N. Weisstein and C.S. Harris, ''Visual detection of line segments: an object-superiority effect'', *Science*, 196, pp. 752–755, 1974.