# THREE LEVELS OF INDUCTIVE INFERENCE

*Peter Gärdenfors*
*Lund University Cognitive Science*
*Kungshuset, Lundagård*
*S–22350 Lund*
*Sweden*
*E-mail: Peter.Gardenfors@fil.lu.se*

## 1. THREE PERSPECTIVES ON OBSERVATIONS

One of the most impressive features of human cognitive processing is our ability to perform inductive inferences. Without any perceived effort, we are prepared, sometimes with great confidence, to generalize from a very limited number of observations.

One of the goals of cognitive science in general, and artificial intelligence in particular, is to provide computational models of different aspects of human cognition. So how can we mechanize induction? How can we even *hope* to capture the ease and assurance of the human inductive competence in a model confined by the thoroughness and strictness of computation?

It is commonplace that induction is going from single observations to generalizations. But this statement loses its air of triviality if one takes seriously, as I propose to do, the question of *what* an observation is. It is surprising that this question has received very little attention within the philosophy of science.[1] The key argument of this article is that there is no unique way of characterizing an observation. Indeed, I shall distinguish three levels of accounting for observations (or, since all levels may be adopted at the same time, they may as well be called perspectives):

*1. The linguistic level:* This way of viewing observations consists of describing them in some specified language. The language is assumed to be equipped with a fixed set of primitive predicates and the denotations of these predicates is taken to be known. As will be argued in Section 2, the linguistic approach is a central part of logical positivism.

*2. The conceptual level:* On this level observations are not defined in relation to some language but characterized in terms of some underlying 'conceptual space'. The conceptual space, which is more or less connected to perceptual mechanisms, consists of a number of 'quality dimensions'. Induction is here seen as closely related to *concept formation*. According to the conceptual perspective, inductive inferences show prototype effects, in contrast to the linguistic perspective which operates on Aristotelian concepts (cf. Smith & Medin 1981).

*3. The subconceptual level:* Observations are here characterized in terms of inputs from sensory receptors. The observations are thus described as occurring before conceptualization. The inductive process is seen as establishing connections between various types of inputs. One currently popular way of modelling this kind of process is by using neural networks.

My main objective in this article is to argue that depending on which approach to observations is adopted, thoroughly different considerations about inductive inferences will come into focus.[2] In my opinion there is a multitude of aspects of inductive reasoning and not something that can be identified as *the* problem of induction. The upshot is that there is no canonical way of studying induction. What is judged to be the salient features of the inductive process depends to a large extent on *what* an observation is considered to be.

[1] One notable exception is Shapere (1982). See Section 4.1.

[2] I cannot talk about three ways of *describing* observations, because the very notion of 'describing' presumes the linguistic level.

Lund University Cognitive Studies – LUCS 9 1992. ISSN 1101–8453.

# 2. THE LINGUISTIC LEVEL

## 2.1 OBSERVATION STATEMENTS AND THE RIDDLES OF INDUCTION

The most ambitious project of analyzing inductive inferences during this century has been that of the logical positivists. According to their program, the basic objects of scientific inquiry are *sentences* or *statements* in some formal or natural language. An *observation* is a particular type of statement. The observational statements are supposed to be furnished to the reasoner by uncorrigible perceptual mechanisms.

Ideally, the scientific language is a version of first order logic where a designated subset of the atomic predicates represent *observational* properties and relations. These observational predicates are taken to be *primitive* notions. This means that when it comes to inductive reasoning, all observational predicates are treated in the same way. For example, Carnap (1950, Section 18B) requires that the primitive predicates of a language be logically independent of each other. The advantage of this, from the point of view of the positivists, is that induction then becomes amenable to *logical analysis* which, in the purist form, is the only tool admitted.

However, it became apparent that the methodology of the positivists led to serious problems for their analysis of induction. The most famous ones are Goodman's (1955) "riddle of induction" and Hempel's (1965) "paradox of confirmation". In Gärdenfors (1990), I have analyzed these problems using the conceptual approach to induction. One conclusion to be drawn from the analysis is that these problems show that the linguistic level is not sufficient for a complete understanding of inductive reasoning.

## 2.2 AN EXAMPLE FROM MACHINE LEARNING

The most common type of knowledge representation within the AI tradition is 'propositional' in the sense that it is based on a set of rules or axioms together with a data base. In this representation, the 'facts' in the data base correspond to observations. The rules and the data base are combined with the aid of a theorem prover or some other inference mechanism to produce new rules or facts. The basic and the derived 'knowledge' is then the material on which a planning or problem solving program can operate.

The propositional form of knowledge representation used in mainstream AI is thus well suited to the positivist tradition. And when implementing inductive inference mechanisms on a computer, this has been the dominating methodology. A rather typical example of the linguistic perspective within AI is the chapter on induction in Genesereth and Nilsson (1987). They assume (pp. 161-162) that there is a set $\Gamma$ of sentences which constitutes the *background theory* and a set $\Delta$ of data (which is to be generalized). It is required that $\Gamma$ does not logically imply $\Delta$. They then define a sentence $\phi$ to be an *inductive conclusion* if and only if (1) $\phi$ is consistent with $\Gamma \cup \Delta$ and (2) the hypothesis $\phi$ *explains* the data in the sense that $\Gamma \cup \{\phi\}$ logically entails $\Delta$.[3]

In general, Genesereth and Nilsson view inductive inferences as problems of *concept formation*:[4]

> The data assert a common property of some objects and deny that property to others, and the inductive hypothesis is a universally quantified sentence that summarizes the conditions under which an object has that property. In such cases, the problem of induction reduces to that of forming the *concept* of all objects that have that property. (1987, p. 165).

They define a *concept-formation problem* as a quadruple $\langle P,N,C,\Lambda\rangle$, where $P$ is a set of positive instances of a concept, $N$ is a set of negative instances, $C$ is a set of concepts to be used in defining the concept, and $\Lambda$ is a language to use in phrasing the definition.

For example, consider the problem of identifying a class of cards from a regular card deck. The language for problems of this kind of problem is taken to be a standard first order language with a set of basic predicates like 'numbered', 'face', 'odd', 'jack', 'four', 'red', and 'spade'. The set $P$ consists of those cards we know belong to the class and $N$ consists of the cards we know are not members of the class. The 'conceptual bias' $C$ determines which among the basic predicates are allowed to be used in forming the inductive rule determining the class. For example, only 'numbered', 'face', 'black' and 'red' may be allowed when describing the rule, so that 'bent' and 'played with the left hand', among others, are excluded. $\Lambda$, finally, is the 'logical bias' which restricts the logical form of the rule that determines the class. For instance, only definitions consisting of conjunctions of basic predicates may be allowed.

[3] Note that this criterion can only be seen as supplying necessary but not sufficient conditions. For example, for any sentence $\alpha$ such that $\Gamma$ logically entails $\alpha$, it holds that $\neg\alpha \vee \bigwedge\Delta$ (where $\bigwedge\Delta$ is the conjunction of all elements in $\Delta$) is consistent with $\Gamma \cup \Delta$ and $\Gamma \cup \{\neg\alpha \vee \bigwedge\Delta\}$ logically entails $\Delta$.

[4] For a similar approach, see Michalski and Stepp (1983).

Using the notion of a concept-formation problem $\langle P,N,C,\Lambda\rangle$, Genesereth and Nilsson develop an algorithm for performing inductive inferences satisfying the constraints given by $C$ and $\Lambda$. A central notion in their construction is that of the 'version space' for the concept-formation problem which consists of all rules that are satisfied by all the positive instances in $P$, but by no instance in $N$. The algorithm works by pruning the version space as new positive and negative instances are added.

Even though AI researchers have had some success in their attempts to mechanize induction, it is clear that their methodology suffers from the same general problems as the linguistic level in general. The enigmas of induction that have been unearthed by Goodman, Hempel and others are applicable also to the induction programs in recent mainstream AI.

Trying to capture inductive inferences by an algorithm also highlights some of the general limitations of the linguistic perspective. The programs work by considering the applicability of various logical combinations of the atomic predicates. But the epistemological origin of these predicates are never discussed. Even though AI researchers are not actively defending the positivist methodology, they are following it implicitly by treating certain predicates as observationally, or at least externally, given. However, the fact that the atomic predicates are assumed as granted from the beginning means that much inductive processing has already been performed.

I agree with Genesereth and Nilsson (1987) that induction is *one form* of concept formation, but their sense of concept formation is *much too narrow*. We not only want to know how observational predicates should be combined in the light of inductive evidence, but, much more importantly, *how the basic predicates are inductively established* in the first place. This problem has, more or less, been swept under the rug by the logical positivists and their programming followers in the current AI tradition. Using logical analysis, the prime tool of positivism and AI, is of no avail for these forms of concept formation. In brief, the linguistic approach to induction sustains no creative inductions, no genuinely new knowledge, and no conceptual discoveries. To do this, we have to go below language.

# 3. The conceptual level

What I see as the source of the troublesome cases for the linguistic approach, like Hempel's and Goodman's riddles, is that if we use *logical* relations alone to determine which inductions are valid, the fact that all predicates are treated on a par induces *symmetries* which are not preserved by our understanding of the inductions: "Raven" is treated on an equal basis with "non-raven", "green" with "grue" etc. What we need is a non-logical way of distinguishing those predicates that may be used in inductive inferences from those that may not.

There are several suggestions for such a distinction in the literature. One idea is that some predicates denote "natural kinds" or "natural properties" while others don't, and it is only the former that may be used in inductions (cf. Quine 1969 and Gärdenfors 1990). Natural kinds are normally interpreted realistically, following the Aristotelian tradition, and thus assumed to represent something that exists in reality independently of human cognition. However, when it comes to inductive inferences it is not sufficient that the properties exist out there somewhere, but we need to be able to *grasp* the natural kinds with our minds. In other words, what is required to understand induction, as performed by humans, is a conceptualistic or cognitive analysis of *observations* of natural properties. Thus we are back at the problem of saying what an observation is, but now on the conceptual level.

## 3.1 Conceptual spaces

One of the primary functions of concepts is to structure the perceptual sensory inflow into categories that are useful for planning, reasoning and other cognitive activities. The concepts we use are not independent of each other but can be structured into *domains*: Spatial concepts belong to one domain, kinship relations to another, concepts for sounds to a third, and so on.[5]

The epistemological framework for a domain of concepts I propose to call a *conceptual space*. A conceptual space consists of a number of *quality dimensions*. I have no exhaustive definition of a what a quality dimension is, but must confine myself to giving examples. Some of the dimensions are closely related to what is produced by our sensory receptors like space, pitch, temperature and color, but there are also quality dimensions that are of an abstract non-sensory character like time and dimensions of social relations. The dimensions of a conceptual space are taken to be cognitive and infra-linguistic in the sense that we (and other animals) can represent the properties of objects, for example when planning an action, without presuming an internal language in which these properties are expressed.

The notion of 'space' should be taken in the mathematical sense. It is assumed that each of the quality dimensions is endowed with certain *topological* or *metric* structures. For example, 'time' is a one-dimensional structure which we conceive of

[5] Cf. Langacker's (1986) use of 'domains' in cognitive semantics.

as being isomorphic to the line of real numbers.[6] Similarly, 'weight' is one-dimensional with a zero point, isomorphic to the half-line of non-negative numbers. The topological structure of the color space is described in Gärdenfors (1990). Some quality dimensions have a discrete structure, i.e., they merely divide objects into classes, e.g., the sex of an individual.[7]

Let us now turn to the problem of identifying observations on the conceptual level. Using the notion of conceptual spaces, an observation can be defined as *an assignment to an object of a location in a conceptual space*. For example, the observation that is described on the linguistic level as "x is red" is expressed on the conceptual level by assigning x a point in color space. Since natural languages only divide the color domain into a finite number of categories the information contained in the statements that x is red is much less precise than the information furnished by assigning x a location in color space. In this sense, the conceptual level allows much richer devices for reporting observations.

## 3.2 CONCEPT FORMATION

On the conceptual level one can distinguish between two types of inductive processes. One is closely related to *concept formation*: In Gärdenfors (1990), I analysed 'natural properties' in terms of conceptual spaces. The key idea is that a natural property is identified with a *convex region* of a given conceptual space. Via the notion of 'convexity' the topological properties of the quality dimensions are utilized. A convex region is characterized by the criterion that for every pair $o_1$ and $o_2$ of points in the region, all points between $o_1$ and $o_2$ are also in the region. The definition presumes that the notion of 'between' is meaningful for the relevant dimensions. This is, however, a rather weak assumption which demands very little of the underlying topological structure.

On the basis of this criterion of natural properties, it is now possible to formulate a constraint on induction, which is helpful in solving the conundrums of the linguistic approach:

*(C) Only properties corresponding to a convex region of the underlying conceptual space may be used in inductive inferences.*

It is only proposed that convexity is a necessary condition, but perhaps not sufficient, for a property to count as natural and thus allowed in inductive inferences. I argue in Gärdenfors (1990) that criterion (C) solves many of the problems of induction that appear on the linguistic level. Furthermore, the criterion can also be used to explain the prototype effects that are exhibited by natural concepts (Rosch 1975, 1978, Gärdenfors 1991).

An assumption that is within reach now is that most basic words in natural languages denote convex regions in some conceptual space. (This assumption can be made even if we have no idea of what the dimensions are or how their topology looks like). From the assumption it follows that the assignment of meanings to the expressions on the linguistic level is far from arbitrary. On the contrary, the semantics (and to some extent even the grammar) of the linguistic constituents is severely constrained by the structure of the underlying conceptual space. This thesis is anathema for the Chomskian tradition within linguistics, but, as a matter of fact, it is one of the central tenets of the recently developed 'cognitive' linguistics.[8]

As another sign of the importance of the conceptual level, I submit that most of scientific theorizing takes place at this level. Determining the relevant dimensions involved in the explanation of a phenomenon is a prime scientific activity. And once the conceptual space for a theory has been established, theories, in the form of *equations*, that connect the dimensions can be proposed and tested.[9]

## 3.3 THE ORIGIN OF QUALITY DIMENSIONS

The second kind of inductive process on the conceptual level concerns how the quality dimensions of the conceptual spaces are determined. There does not seem to be a unique origin of our quality dimensions. Some of the dimensions are presumably *innate* and to some extent hardwired in our nervous system, as for example color, pitch, and probably also ordinary space. These subspaces are obviously extremely important for basic activities like finding food and getting around in the environment.

But from the point of view of induction, the dimensions that are *learned* are of greater interest. Learning new concepts often involves expanding one's conceptual space with new quality dimensions. 'Volume' is an example here. According to Piaget's 'conservation' experiments with five year olds, small

[6] To some extent the representation of time is culturally dependent, so that other cultures have a different time dimension as a part of their cognitive structure. Cf. Gärdenfors (1992) for a discussion of how this influences the structure of language.

[7] Discrete dimensions may also have additional structure as, for example, in kinship or biological classifications. The topology of discrete dimensions is further discussed in Gärdenfors (1990).

[8] Cf. Lakoff (1987) and Langacker (1986).

[9] For a discussion of the role of conceptual spaces in science, see Gärdenfors (1990) and (1991).

children do not make a distinction between the height of a liquid and its volume. The conservation of volume, which is part of its conceptual structure, is something that must be learned. In general, introducing *new* quality dimensions is a much more advanced form of induction than concept formation *within* a given conceptual space.

A similar process occurs within *science*. By introducing theoretically precise, non-psychological quality dimensions, a scientific theory may help us find new inductive inferences that would not be possible on the basis of our subjective conceptual spaces alone. As an example, consider Newton's distinction between weight and mass, which is of crucial importance for the development of his celestial mechanics, but which has no correspondence in human psychology. It seems to me that the cognitive construction involved in Newton's discovery of the distinction between mass and weight is of the same nature as when a child discovers the distinction between height and volume. Another example of a scientifically introduced dimension is the distinction between temperature and heat, which is central for thermodynamics. In contrast, human perception of heat is basically determined by the amount of heat transferred from an object to the skin rather than by the temperature of the object.

In order to give another illustration of how the scientific process is helpful in constructing the underlying conceptual space, thereby providing an understanding of how concepts are formed, I shall briefly present the phonetic identification of *vowels* in various languages. According to phonetic theory, what determines a vowel are the relations between the basic frequency $F_0$ of the sound and its formants (higher frequencies that are present at the same time). In general, the first two formants $F_1$ and $F_2$ are sufficient to identify a vowel. This means that the coordinates of two-dimensional space spanned by $F_1$ and $F_2$ (in relation to a fixed basic pitch $F_0$) can be used as a fairly accurate description of a vowel. Fairbanks and Grubb (1961) investigated how people produce and recognize vowels in 'General American' speech. Figure 1 summarizes some of their findings.

The scales of the abscissa and ordinate are the logarithms of the frequencies of $F_1$ and $F_2$ (the basic frequency of the vowels was 130 cps). A *self-approved* vowel is one that was produced by the speaker and later approved of as an example of the intended kind.

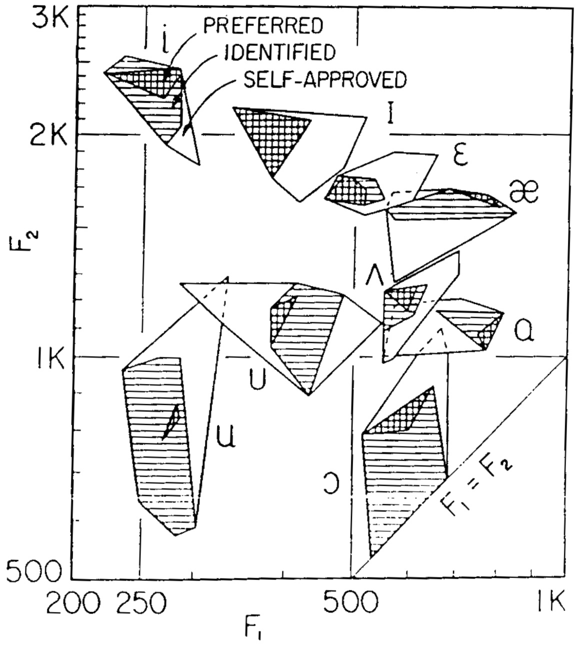

Figure 1.
Frequency areas of different vowels in the two-dimensional space generated by the first two formants. Values in cps. (From Fairbanks and Grubb, 1961.)

An *identified* sample of a vowel is one that was correctly identified by 75% of the observers. The *preferred* samples of a vowel are those which are "the most representative samples from among the most readily identified samples" (Fairbanks and Grubb 1961, p. 210).

As can be seen from the diagram, the preferred, identified and self-approved examples of different vowels form convex subregions of the space determined by $F_1$ and $F_2$ with the given scales. As in the case of color terms, different languages carve up the phonetic space in different ways (the number of vowels identified in different languages varies considerably), but I conjecture again that each vowel in a language will correspond to a convex region of the formant space.

The important thing to note in this example is that identifying $F_1$ and $F_2$ as the relevant dimensions for vowel formation is a phonetic *discovery*. We had the concepts of vowels already before this discovery, but the spatial analysis makes it possible for us to understand several features of the classifications of vowels in different languages.

The last examples of how science introduces new quality dimensions for concept formation highlights one fundamental problem for this second type of inductive process on the conceptual level: *Where do the dimensions and their topology come from?* According to Popper's terminology this kind of process belongs to the 'context of discovery'. Within traditional philosophy of science, it has in general been thought to be futile to construct a mechanistic procedure for generating scientific discoveries of this kind. However, when it comes to human learning and concept formation, the prospects may not be so hopeless after all. This will be the topic of next section where inductive processes below the conceptual level will be considered.

# 4. The subconceptual level

## 4.1 Observations by receptors

In the most basic sense an observation is what is received by our sensory organs. In this sense, an observation can be identified with what is received by a set of *receptors*. For human beings, these inputs are provided by the sensory receptors, but one can also talk of a machine having observations of this kind via some measuring instruments serving as receptors. The receptors provide 'raw' data in the sense that the information is not assumed to be processed in any way, neither in a conceptual space, nor in the form of some linguistic expression.

Within the philosophy of science, it is important to make a distinction between *perception* and *observation*. As Shapere (1982) points out, the term 'observation' plays a double role for the traditional philosopher of science. He writes:

> On the one hand, there is the *perceptual* aspect: "observation", as a multitude of philosophical analyses insist, is simply a special kind of perception, usually interpreted as consisting in the addition to the latter of an extra ingredient of focussed attention. ... On the other hand, there is the *epistemic* aspect of the philosopher's use of 'observation': the *evidential* role that observation is suppose to play in leading to knowledge or well-grounded belief or in supporting beliefs already attained. (Shapere 1982: 507-508)

Within the empiricist tradition of philosophy of science, the two uses of 'observation' have been confounded. However, in modern science it is obvious that it is the epistemic aspect of observation that is of importance. As Shapere (1982: 508) formulates it:

> Science is, after all, concerned with the role of observation as evidence, whereas sense-perception is notoriously untrustworthy ... . Hence, with the recognition that information can be received which is not directly accessible to the senses, *science has come more and more to exclude sense-perception as much as possible from playing a role in the acquisition of observational evidence;* that is, it relies more and more on other appropriate, but dependable, receptors.

Given that we are focussing on the epistemic aspect of observations, let us then consider induction on the subconceptual level. How do we distill sensible information from what is received by a set of receptors? Or, in other words, how do we make the transition from the subconceptual to the conceptual and the linguistic levels? These questions indicate the kinds of inductive problems that occur on the subconceptual level.

The basic problem is that the information received by the receptors is too rich and unstructured. What is needed is some way of transforming and organizing the input into a form that can be handled on the conceptual or linguistic level. There are several methods for treating this kind of problem. Within psychology, various methods of *multidimensional scaling* have been developed.

For example, in Shepard's (1962a,b) algorithm, the input data is assumed to contain information about the relative distances between $n$ points in some unknown space. The distances between the points are not expressed in metrical terms, but only given as a rank order of the $n(n-1)/2$ distances between the $n$ points. Any such rank order can be represented in a space of $n$-1 dimensions. Shephard's algorithms starts out from a representation in such a space and then successively reduces the dimensionality until no further dimensions can be eliminated without a substantial disagreement between the rank order generated by the metric assignment and the original rank order. For many empirical areas the intial data can be reduced to a space with two or three dimensions.[10] These dimensions can then function as a basis for concept formation according to the outline in Section 3.2.

## 4.2 Induction with the aid of neural networks

In this subsection a different method for going from the subconceptual to the conceptual level will be outlined. The mechanisms for the method are based on *neural networks*. In a neural network, the receptors and the information they receive can be identified with a set of *input neurons* and their *activity values*. This set of values will be called the *input vector*. In interesting cases there is a large number of input neurons which means that the dimensionality of the

[10] Cf. Shepard (1962b) for several examples of the results of the procedure.

input vector is very high. The purpose of an inductive method at this subconceptual level is to reduce the complexity of the input information in an efficient and systematic way.

The neural network model I will be outlining here is based on Kohonen's (1988) *self-organizing feature maps*. The distinguishing property of these maps is that they are able to describe the topological relations of the signals in the input vector using something like a conceptual space with a small number of dimensions. Basically, the mapping can be seen as reducing the dimensionality of the input vector.

A self-organizing feature map is a neural network which consists of an input vector that is connected to an output array of neurons. In most applications, this array is one- or two-dimensional, but in principle it could be of any number of dimensions. The essential property of the network is that the connections between the neurons in the array and the learning function are organized in such a way that *similarities* that occur among different input vectors are *preserved* in the mapping, in the sense that input vectors that have common features are mapped onto *neighbouring* neurons in the map. The degree of similarity between two input vectors is determined by some *distance* measure (which normally is the standard Euclidean metric, but many metrics are possible to use).

In other words, the mapping from the input vector to the array preserves the topological relations while reducing the dimensionality of the representation space. The low-dimensional 'feature map' that results as an output of the process can be viewed as a conceptual space in the sense of the preceeding section. The mapping is *generated* by the network itself via the learning mechanism of the network. In practice, it normally takes quite a large number of learning instances before the network stabilizes enough so that further changes can be ignored.[11]

The mechanism is best illustrated by a couple of artificial examples taken from Kohonen (1988). In figures 2 and 3 the input vectors were assumed to be uniformly distributed over a triangular area. In the network represented in figure 2, the output array was one dimensional, i.e., the output neurons were arranged along a line. The number of neurons on this line is fixed. Any input from the triangular space results in some activities in the neurons in this line. Figure 2 shows the inverse mapping of the input vectors which resulted in the highest activities of single neurons in the line, where each dot corresponds to an output neuron. As can be seen, the mapping preserves relations of similarity, and, furthermore, there is a tendency of the line trying to 'cover' as much as possible of the surface, in the sense that the distance between any point in the surface and the line being as small as possible.

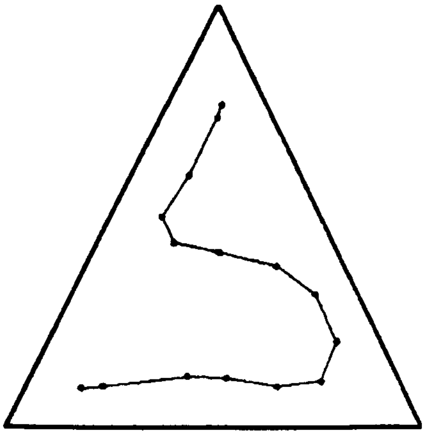

Figure 2.
Distribution of weight vectors on a linear array of neurons.
(From Kohonen (1988), p. 135.)

In figure 3, the corresponding network contains an output array that is two-dimensional, with the neurons arranged in a square. Figure 3 again shows the inverse mapping, indicating which neurons in the input space produce the greatest responses in the output square. As can be seen, the inverse mapping represents a deformation of the output array that preserves topological relations as much as possible.

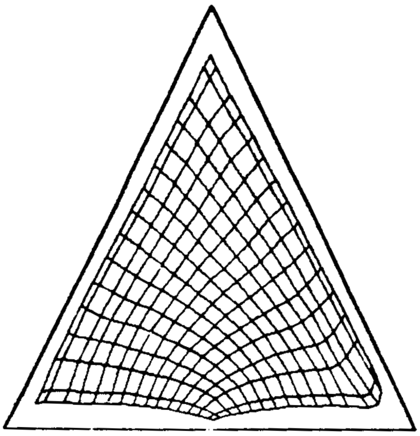

Figure 3.
Distribution of weight vectors on a rectangular array of neurons.
(From Kohonen (1988), p. 135.)

[11] New learning by instances that do not follow the previous frequency pattern can always change the mapping function. This means that it is impossible to talk about a 'final' mapping function.

Figure 4 shows an example of how the network self-organizes in learning a mapping. The initial values of the mapping were selected so that there was a random mapping from a circular region of the input triangle to a linear array of output neurons. The network was then fed with a number of input vectors, randomly selected from the full triangle. The sequence of figures indicate how the mapping is improved over time, where the numbers below the figures represent the number of learning trials.

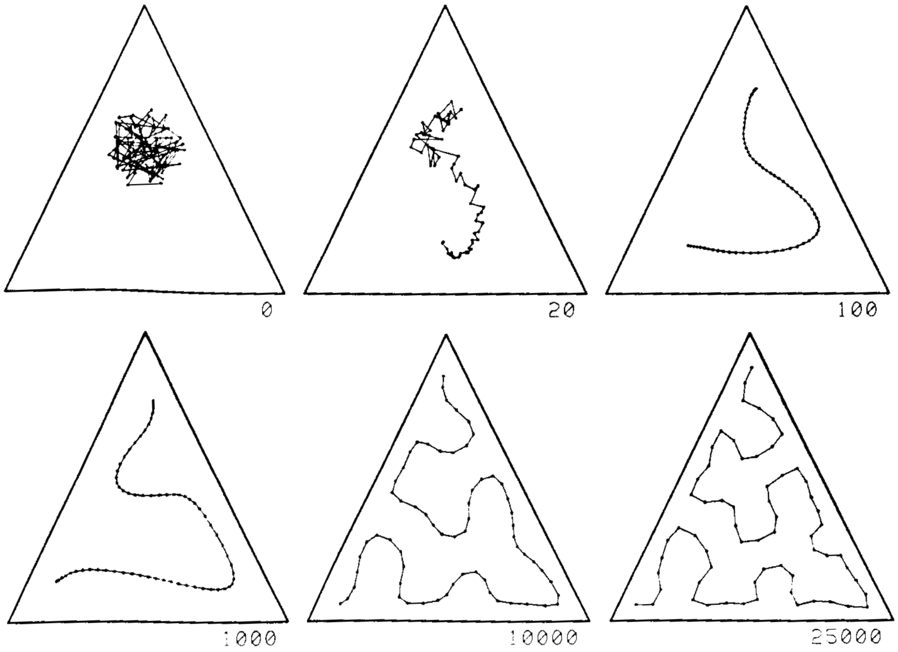

Figure 4.
Distribution of weight vectors on a rectangular array of neurons.
(From Kohonen (1988), p. 135.)

These examples are artificial in that we know the initial distribution of input vectors, which furthermore is of low dimensionality. In real applications, the dimensionality of the input space is high and its topology is unknown. However, it can be shown, at least when the output array is one-dimensional, that the mapping in the limit (i.e., after infinitely many learning instances) will preserve as much as possible of the topological structure of the input space.[12]

Kohonen's goal in using the maps is not limited to inductive inference only but representation of information in general. He writes:

> Economic representation of data with all their interrelationsships is one of the most central problems in information sciences, and such an ability is obviously characteristic of the operation of the brain, too. In thinking, and in the subconscious information processing, there is a general tendency to compress information by forming *reduced representations* of the most rele-vant facts, without loss of knowledge about their interrelationsships (Kohonen 1988, p. 119).

In collaboration with Christian Balkenius, I have started working on a general architecture for a neural network system that utilizes self-organizing feature maps to perform inductive inferences. The overall architecture of the inductive network is depicted in Figure 5. The input receptors are divided into a small number of subsets (in the figure there are two such subsets). The purpose of this division is to group together receptors that contain information about 'the same' feature, so for example, visual receptors belong to one group, auditory receptors to another etc. When the network is applied, the decision about how the set of receptors should be grouped must be made by the user. But this is about the only thing she has to decide – except for some parameter settings, the network then performs the rest of the inductive inferences.

The input vectors are then mapped onto one Kohonen surface each. In figure 5 these are depicted as one-dimensional lines, but they may as well be two- or three-dimensional surfaces. In the figure, there are

[12] For a more precise statement of this result and a proof see Kohonen (1988), pp. 145-148.

only two Kohonen surfaces, but they may, of course, be more than two depending on how the input receptors are grouped into subspaces. One of the surfaces may be a purely classificatory space, representing 'names' of the categories that are identified by the network.[13]

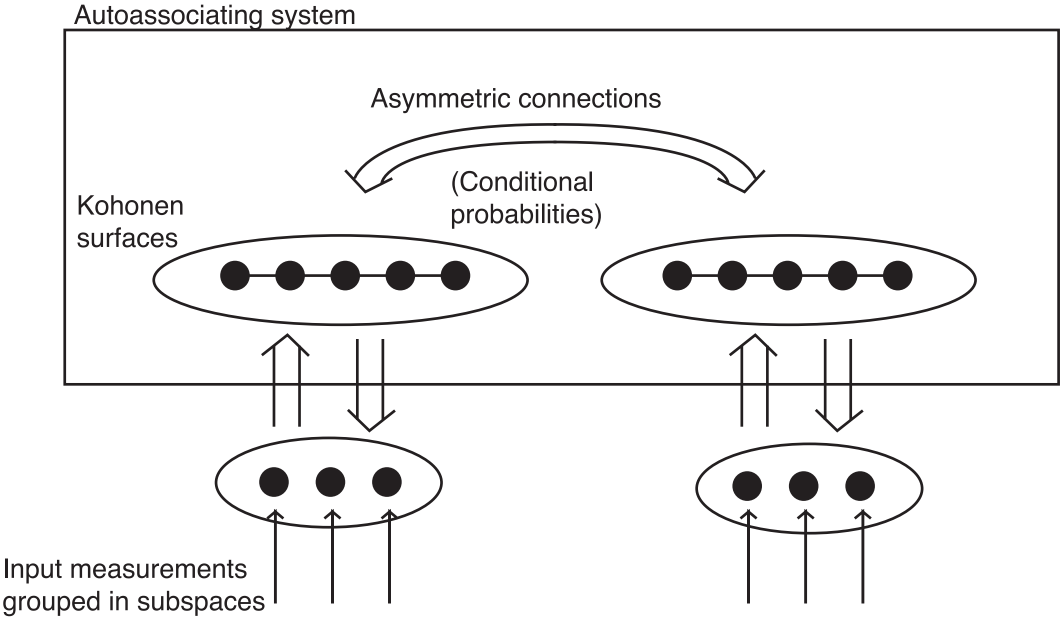
Autoassociating system
Asymmetric connections
(Conditional probabilities)
Kohonen surfaces
Input measurements grouped in subspaces

Figure 5.
Architecture of inductive neural network

The Kohonen surfaces are then pairwise connected by asymmetric connections between the neurons in the two surfaces. The connections are total in the sense that each neuron on one surface is connected to all neurons on the other surface. The learning rule for these connections functions in such a way that the strength of the connection $c_{ij}$ between a neuron $x_i$ on one surface and a neuron $y_j$ on another reflects the conditional probability (estimated from the learning examples) that $y_j$ be activated given that $x_i$ is activated.[14] The connections vary between -1 and +1 and obtain the extreme values only when $x_i$ and $y_j$ are never and always, respectively, activated together. In a sense, the network performs *implicit* computations of the inductive statistics.[15]

Once the learning phase has been completed, relating input receptors to Kohonen surfaces and these surfaces to each other, it is then possible to use the network to classify new objects. By feeding the system with a *partial* input vector, for example, the values for one of the subspaces, the network can then compute the *expected* values for all the other receptors and the expected locations for all the Kohonen surfaces. In this way the network *guesses* the unknown properties of the object it has only received partial information about. The network is thus able to *generalize* from earlier experience and make inductive inferences using the connections between the different Kohonen spaces.

Christian Balkenius and I have done some preliminary experiments using a network with the architecture that has been outlined here. So far, the results seem very promising. One example concerns a classification of 44 individual parasitical wasps. For each individual the values of twelve variables are supplied together with the species name it was assigned by an entomologist. These variables represent different kinds of data, some binary, some ordinal, and some real-valued. After discussions with the entomologist, we divided the input variables into four groups: One consisting of five variables on proportions of sizes of various body parts, the second consisting of four other morphological variables, the third consisting of three ecological variables, and the fourth simply a coding for the species name. Each of these four variable groups was mapped onto a one-dimensional Kohonen surface (i.e., a line), and the four surfaces were pairwise connected by asymmetric connections as described above.

[13] The linguistic form of the names has, of course, to be provided by the user.

[14] The mathematical form of the connections are closely related to Hintikka's (1969, p. 328) measures of 'evidential support,' in particular the measure defined in his equation (27)*.

[15] The sense in which neural networks perform implicit statistic inferences can be made very precise. For example, see Lippman (1987) for a presentation of some of the results connecting least mean square and maximal likelihood solutions to the computations of neural networks.

After training the network by showing it the individual input vectors a number of times, it can be tested by feeding in all input variables for a particular individual, except for its species categorization and compare the output with that of the entomologist. In our tests, the network makes very few errors in classifying the 44 wasps. The results are, to some extent, dependent on the number of neurons on each Kohonen surface. If we allow as many as 50 neurons, which is more than the number of wasps, then the network can learn to correctly classify every individual in the sample. However, if there are only 20 neurons on the Kohonen surfaces, then it correctly classifies 43 out of 44. One can, of course, also feed in data (except for names of the species) about individuals that were not present in the learning sample. The species names that are produced as outputs from the network seem to have a high degree of validity. However, the performance of the network still awaits more detailed testing against empirical material.

The methodology of founding a classification on a large number of numerical data is similar to so called numerical (or phenetic) taxonomy (Sokal and Sneath 1963, Ruse 1973) In my opinion, however, the mechanisms behind the classifications obtained by the neural network model and their biological validity are superior to what is achieved in numerical taxonomy. One essential difference is that, as a matter of methodology, all variables are treated as having equal value in the process of computing the numerical taxonomy. However, even if it may seem that the same holds of the inputs to the neural network described above, the variables are *implicitly* assigned different importance via the influence they have on the topology of the Kohonen surfaces that emerge during the learning period.

What are then the drawbacks of using neural networks of the type described here for inductive processes? A fundamental epistemological problem is that even if we know that the network will generate Kohonen surfaces that perform the right kind of job, we may not be able to 'describe' *what* the emerging dimensions represent. Even if we, for example, know that a system consisting of four one-dimensional Kohonen surfaces provides a perfect classification of a population of parasitical wasps, this may not help us in *interpreting* the 'meaning' of the surfaces, i.e., what overall features of the wasps that they *represent*. In other words, we may not be able to make the transition between the subconceptual level and the conceptual level. This kind of level problem is ubiquitous in applications of neural networks for learning purposes. The upshot is that a future theory of neural networks must somehow bridge bridge the gap of going from the subconceptual level to the conceptual level. We may account for the information provided at the subconceptual level in term of a dimensional space with some topological structure, but there is no general recipe for determining what is the *conceptual* meaning of the dimensions of the space.

Other problems concern more methodological issues. How should the input variables be grouped before they are mapped onto different Kohonen surfaces? How does one decide how many dimensions to use in the target surface of the mapping from the input variables? These kinds of problems are found everywhere in science and they are not particular to using neural networks for the classifications.[16] In fact, the methodological problems involved in the procedure presented here seem to be smaller than the problems one encounters for other classification methods.

## 5. Conclusion: What is induction?

Where on the three levels that have been described here is real induction to be found? The answer is nowhere and everywhere. The main thesis of this article is that there are several kinds of inductive processes. Depending on what perspective one takes on observations, different ways of generalizing the observations become relevant. Traditional philosophy of science has concealed these distinctions by neglecting the conceptual and subconceptual levels. For a complete account of induction, all three levels must be mustered.

What is the relation between the three levels? I hope it has become clear from my presentation that I do not view the three levels as being in conflict with each other. They should rather be regarded as three *perspectives* on observations that complement each other. Different aspects of inductive processes need to be explained on different levels. By disregarding some level one restricts the possibilities for understanding the mechanisms of inductive reasoning.

A three-level theory of cognitive representation that is related to the one proposed in this paper has been suggested by Harnad (1987) as a way of analysing problems in categorical perception.[17] He calls his lowest level the *iconic* representation (IR), "being an analog of the sensory input (more specifically, of the proximal projection of the distal stimulus object on the device's transducer surfaces)" (Harnad 1987, p. 551). The IRs are analog mappings which "faithfully

[16] For instance, very similar problems would be encountered when applying the multi-dimensional scaling methods that were outlined above.

[17] This theory was brought to my attention by Paul Hemeren when the present article was almost finished.

preserve the iconic character of the input for such purposes as same-different judgements, stimulus-matching, and copying" (p. 552). It is obvious that this form of representation corresponds to what I have here called the sub-conceptual level.

The middle level Harnad calls *categorical* representation (CR). This representation eliminates most of the raw input structure and retains what is invariant in the produced categorization: "Whereas IRs preserve analog structure relatively indiscriminately, CRs selectively reduce input structure to those invariant features that are sufficient to subserve successful categorization (in a given context)" (p. 553).

Again, it is clear that this level corresponds to the conceptual level of this article. Unfortunately, Harnad says very little about how the categorization is acheived, except that it is some kind of filtering process. Furthermore, he provides no account of the structure of the categorical representation, with the exception that he presumes that categorization is to a certain extent context dependent. I believe that it is a strength of the theory of conceptual spaces outlined in Section 3 that it has strong, and to a large extent testable, implications for categorization and concept formation.

The highest level in Harnad's triad is *symbolic* representation (SR), which naturally corresponds to the linguistic level of this paper. He introduces a "description system" (p. 554), the expressions of which assign category membership to experiences. The description system presumes that the CRs are already *labeled*:

> Instead of constructing an invariance filter of the basis of direct experience with instances, it operates on *existing* labels, and constructs categories by manipulating these labels, in particular, assigning membership on the basis of stipulated rules rather than perceptual invariants derived from direct experience (p. 554).

Here it seems to me that Harnad is partly falling back on the Aristotelean tradition of concept formation. The upshot seems to be a hybrid theory:

> Descriptions spare us the need for laborious learning by direct acquaintance; however, they depend on the prior existence of a repertoire of labeled categories on which the combinatory descriptions can draw. Hence *symbolic* representations (SRs), which are encoded as mental sentences, define new categories, but they must be grounded in old ones; the descriptive system as a whole must accordingly be *grounded* in the acquaintance system (p. 556).

The use of the metaphor "grounded" indicates that Harnad views the three representation forms as separate systems. In contrast, the three levels presented here are three *perspectives* on one and the same system. Nevertheless, the similarities between mine and Harnad's are indisputable. Since Harnad proposes his three kinds of representations as a tool for understanding phenomena of categorical perception, these similarities strengthen the links between concept formation and the present analysis of induction.

It is also worthwhile comparing the three levels of observation and induction discussed in this article with Smolensky's (1988) distinction between the subsymbolic and symbolic levels in the context of connectionist models. In my opinion, his 'subsymbolic level' corresponds closely enough to what has here been called the subconceptual level. However, Smolensky confounds the symbolic and conceptual levels.[18] The reason why is simple: he is committing himself to 'High Church Computationalism' by "limiting consideration to the Newell/Simon/Fodor/Pylyshyn view of cognition" (p. 3). One of the central tenets of the symbolic approach is what Smolensky formulates as 'hypothesis 4b':

> The programs running on the intuitive processor are composed of elements, that is, symbols, referring to essentially the same concepts as the ones used to consciously conceptualize the task domain (p.5).

He then gives the following reason for calling the symbolic level 'conceptual':

> Cognitive models of both conscious rule application and intuitive processing have been programs constructed of entities which are *symbols* both in the syntactic sense of being operated on by symbol manipulation and in the semantic sense of (4b). Because these symbols have the conceptual semantics of (4b), I am calling the level of analysis at which these programs provide cognitive models the *conceptual level* (*ibid.*).

However, there is a different tradition within cognitive science where the conceptual level of this paper is given independent standing. For example, I believe the theory of conceptual spaces presented in Section 3.1 can be seen as a generalization of the *state space* approach, advocated among others by P. M.

[18] He even uses the two names: "I will call the preferred level of the symbolic paradigm the *conceptual* level and that of the subsymbolic paradigm the *subconceptual* level" (Smolensky 1988:3).

Churchland (1986a,b), and of the *vector function theories* of Foss (1988). The theory of conceptual spaces is *a theory for representing information*, not a theory about symbol manipulation. (The symbol paradigm that Smolensky is referring to is called the ‘sentential paradigm’ by the Churchlands.[19])

Even though he fails to identify it as a separate level, Smolensky is well aware of this ‘vectorial’ approach, as can be seen from the following quotation:

> Substantive progress in subsymbolic cognitive science requires that systematic commitments be made to vectorial representations for individual cognitive domains. [...] Unlike symbolic tokens, these vectors lie in a topological space in which some are close together and others far apart (Smolensky 1988, p. 8).

He even recognizes the importance of establishing a connection between the subconceptual level and the conceptual level:

> Powerful mathematical tools are needed for relating the overall behavior of the network to the choice of representational vectors; ideally, these tools should allow us to *invert* the mapping from representations to behavior so that by starting with a mass of data on human performance we can turn the mathematical crank and have the representational vectors pop up. An example of this general type of tool is the technique of *multidimensional scaling* (Shephard 1962), which allows data on human judgments of similarity between pairs of items in some set to be tuned to vectors for representing those items (in a sense). The subsymbolic paradigm needs tools such as a version of multidimensional scaling based on a connectionist model of the process of producing similarity judgments (*ibid.)*

In conclusion, Smolensky’s binary distinction between the symbolic and the subsymbolic level is insufficient. We need all three levels of representing information that have been presented in this paper to give an adequate description of the various inductive processes that are encountered in the human realm as well as in the artificial.

## Acknowledgements

I wish to thank Christian Balkenius, Paul Davidsson, Jens Erik Fenstad, Paul Hemeren, Patrick Suppes, Peter Williams, and the Cognitive Science group in Lund for helpful discussions. My work with this article has been supported by the Swedish Council for Research in the Humanities and Social Sciences.

[19] Cf. P. S. Churchland (1986) and Gärdenfors (1992) for a discussion of the conflict between the two approaches.

# References

Carnap, R. (1950), *Logical Foundations of Probability*, Chicago, IL: Chicago University Press.

Churchland, P. M. (1986a), "Cognitive neurobiology: A computational hypothesis for laminar cortex," *Biology & Philosophy*, **1**, 25-51.

Churchland, P. M. (1986b), "Some reductive strategies in cognitive neurobiology," *Mind*, **95**, no. 379, 279-309.

Churchland, P. S. (1986), *Neurophilosophy: Toward a Unified Science of the Mind/Brain*, Cambridge, MA: Bradford Books, MIT Press.

Fairbanks, G. and Grubb, P. (1961), "A psychophysical investigation of vowel formants," *Journal of Speech and Hearing Research*, **4,** 203-219.

Foss, J. (1988), "The percept and vector function theories of the brain," *Philosophy of Science*, **55**, 511-537.

Gärdenfors, P. (1990), "Induction, conceptual spaces and AI," *Philosophy of Science*, **57**, 78-95.

Gärdenfors, P. (1991), "Frameworks for properties: Possible worlds vs. conceptual spaces," *Acta Philosophica Fennica* **49**, 383-407.

Gärdenfors, P. (1992), "Mental representation, conceptual spaces, and metaphors," forthcoming in *Synthese*.

Genesereth, M. and Nilsson, N. J. (1987), *Logical Foundations of Artificial Intelligence*, Los Altos, CA: Morgan Kaufmann.

Goodman, N. (1955), *Fact, Fiction, and Forecast*, Cambridge, MA: Harvard University Press.

Harnad, S. (1987): "Category induction and representation," in *Categorical Perception*, ed. by S. Harnad, Cambridge: Cambridge University Press, 535-565.

Hempel, C. G. (1965), *Aspects of Scientific Explanation, and Other Essays in the Philosophy of Science*, New York, NY: Free Press.

Hintikka, J. (1969), "The varieties of information and scientific explanation," in *Logic, Methodology, and Philosophy of Science III*, ed. by B. van Rootselaar and J. F. Staal, Amsterdam: North-Holland, 311-331.

Kohonen, T. (1988), *Self-Organization and Associative Memory*, Second Edition, Berlin: Springer-Verlag.

Lakoff G. (1987), *Women, Fire and Dangerous Things*, Chicago, IL: University of Chicago Press.

Langacker, R. (1986), *Foundations of Cognitive Grammar (Vol. 1)*. Stanford, CA: Stanford University Press.

Lippman, R. P. (1987), "An introduction to computing with neural nets," *IEEE ASSP Magazine*, April 1987, 4-22.

Michalski, R. S. and Stepp, R. E. (1983), "Learning from observation: Conceptual clustering," in *Machine Learning, An Artificial Intelligence Approach*, Los Altos, CA: Morgan Kaufmann, 331-363.

Quine, W.V.O. (1969), "Natural kinds," in *Ontological Relativity and Other Essays*, New York, NY: Columbia University Press, 114-138.

Rosch, E. (1975), "Cognitive representations of semantic categories," *Journal of Experimental Psychology: General*, **104,** 192-233.

Rosch, E. (1978), "Prototype classification and logical classification: The two systems," *New Trends in Cognitive Representation: Challenges to Piaget's Theory*, ed. E. Scholnik, Hillsdale, NJ: Lawrence Erlbaum Associates, 73-86.

Ruse, M. (1973), *The Philosophy of Biology*, London: Hutchinson and Co.

Shapere, D. (1982), "The concept of observation in science and philosophy," *Philosophy of Science*, **49**, 485-525.

Shepard, R. N. (1962a), "The analysis of proximities: Multidimensional scaling with an unknown distance function. I.," *Psychometrika*, **27**, 125-140.

Shepard, R. N. (1962b), "The analysis of proximities: Multidimensional scaling with an unknown distance function. II.," *Psychometrika*, **27**, 219-246.

Smith, E. and Medin, D. L. (1981), *Categories and Concepts*, Cambridge, MA: Harvard University Press.

Smolensky, P. (1988), "On the proper treatment of connectionism," *Behavioral and Brain Sciences*, **11**, 1-23.

Sokal, R. R. and Sneath, P. H. A. (1963), *Principles of Numerical Taxonomy*, San Francisco: W.H. Freeman and Co.