Dessalles, J-L. (1998). "Limits of isotropic bias in natural and artificial models of learning". In G. Ritschard, A. Berchtold & F. Duc (Eds), *Apprentissage - Des principes naturels aux méthodes artificielles*. Paris : Hermès, 307-319.

Limits of Isotropic Bias in Natural and Artificial Models of Learning

Jean-Louis Dessalles [1]

Bias is always present in learning systems. There is no perfect, universal, way of learning that would avoid any "innate" predetermination. However, all biases should not be considered equivalent. Usually, it is implicitly regarded as desirable to avoid anisotropic biases when designing a learning mechanism, especially when it is intended as a cognitive model of some human or animal learning ability. Anisotropic bias necessarily involves some *ad hoc a priori* knowledge that severely limits the generality of the learning device.

We want to suggest, however, that isotropic models of learning, though they seem to be of greater generality, may prove to be too limited. In many cases, living beings of the same species reliably learn identical forms from different experiences. We show that these situations, called convergent learning, are hardly explained by isotropic models, unless learned forms are highly harmonious (*i.e.*, symmetrical). This *anisotropy-harmony dilemma* is derived from a formal characterization of bias, based on simple geometrical properties. By showing how this dilemma affects classical theories of learning, we try to clarify the classical nature-nurture debate in the case of convergent learning.

keywords : Learning, isotropy, bias, symmetry, innate knowledge.

1. Introduction

Learning is an adaptive ability of higher animals and of more and more artificial devices. It results in a new ability to discriminate situations that was not present before exposure to experience. We can model this discrimination ability as a classification which, at the end of the learning phase, allows to associate a decision to each situation. Without loss of generality, we will consider a simple device that learns binary classifications (figure 1).

[1] Ecole Nationale Supérieure des Télécommunications - 46 rue Barrault - 75013 PARIS - France - E-mail : dessalles@enst.fr

Theoretical work on Learning Theory is most often concerned with the problem of making good inductions : among a given set of classifications, how to choose a candidate which is good according to a given cost function. For example, the Vapnik law relates the minimum number of examples required by *any* learning device making good inductions to the "separating power" of available classifications [Boucheron 1992].

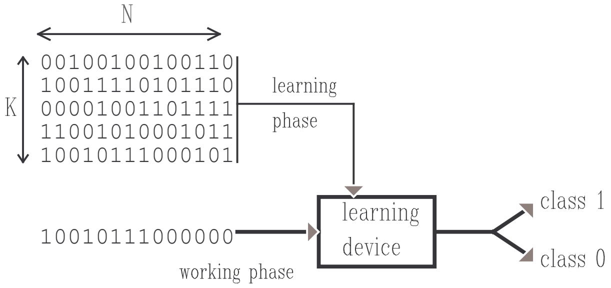

Figure 1 : *A simple learning device. Input is given through N binary sensors. During the learning phase, the system is presented with a sample of K examples of N bits. During the working phase, the system assigns one of two classes (class 1 or class 0) to any binary N-uple.*

In cognitive modeling, however, the learning device is given. From its assumed properties, one draws conclusions upon the set of classifications it may or may not learn. The work presented here belongs to this second approach. It does not refer to cost functions nor to asymptotic correctness. It is based on qualitative, geometrical considerations. Its purpose is to make a link between some intrinsic properties of the learning device (esp. isotropy) and properties of the classifications that this device may reach.

In what follows, we will define such properties : *indifference* for learning devices, *harmony* for classifications. Then we show how both are connected : under certain circumstances, indifferent mechanisms are bound to learn harmonious classifications. We discuss the relevance of this result to cognitive modeling by briefly reviewing some important cognitive learning mechanisms (Gestalt Theory, Piaget's Theory, Associationism, Inneism) and by showing how they comply with these constraints.

2. Bias and specificity of learners

Generalist models of learning are often preferred : as computational learning methods, they are more adaptable, and as cognitive models, they are more parsimonious. Intuitively, generality does not tolerate strong bias. For instance, in [Elman et al., 1997], much effort is devoted to showing that the learning performance of children can be explained without invoking specific innate knowledge about the task. Invoking such knowledge would require that a new learning model has to be postulated for each cognitive competence, whereas in the absence of such specific bias, a single generalist system like connectionism may account for many of the child's abilities.

When deliberately introduced, bias is intended to offer better learning efficiency. However, such improvement will be observed only on a restricted range of situations [Schaffer 1994], and strongly biased systems are thus expected to be more specialized than less biased ones. A generalist learner would ideally rely on virtually no bias. Strictly speaking, this is not possible. Even a simple nearest neighbor device uses a bias : finding the nearest neighbor for a novel datum predetermines the device to one type of inductive generalization instead of another. Can we think of reducing bias to a minimum in order to preserve generality ? The results presented in what follows suggest a negative answer : attempts to avoid unnecessary bias also lead to a certain type of specialization.

Many learning mechanisms that have been proposed, either in cognitive modeling or in computer science, share a property that we call *indifference*. This property results from the avoidance of unnecessary specificity. Most systems claiming to be generalist learners have an indifferent bias. We first define this notion, then we show that this property has interesting consequences which affect what can be learned by such systems.

Let us consider the simple learner of figure 1. Most learning situations can be modeled by such a device. The separation between learning and working phases, the digital representation of input and the restriction to only two classes are not necessary features. The results given below can be extended to devices that lack these limitations (*e.g.* continuous devices).

In the description of figure 1, some components of examples may represent supervision information if any. It is worth noting that we make no assumption upon the ability of the learning device to reach correct or accurate classifications. The learning process is not even supposed to be inductive : the K "examples" are not presumed members of class 0 or 1, and could act as mere triggers in the learning process. In other words, the results presented here do not require the presence of an "oracle" saying whether a learned classification is correct or accurate. Assessing accuracy would be problematic in certain situations encountered in cognitive modeling. What would it mean for human learners that they correctly learned language, word meaning or accent ? In such cases, there is no independent reference telling what is correct and what is not. As we will see, it is nevertheless possible, without considering accuracy, to tell sometimes what a learning system cannot do. We will suggest that some cognitive performances are not the result of isotropic or indifferent learning mechanisms.

3. Isotropy, relativity and indifference of a learning mechanism

"Indifference" is characteristic of devices which do not take *absolute* properties of their input into account. Devices that extract regularities are most often indifferent mechanisms : they only use relative properties of data (distance, sameness). By contrast, a digital sensor that becomes active for a particular configuration of its input is by essence non indifferent : this particular configuration works as an absolute reference. In a learning device, the sensitivity to absolute features hinders the system from learning equivalent forms the same way. This is why indifferent systems are usually preferred : they are more general, they are not constrained by an absolute reference.

The system sketched in figure 2 gives an intuitive idea of what indifference means.

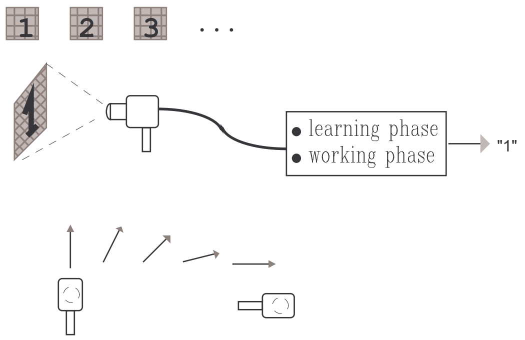

Figure 2 : *illustration of the indifference property. A device using a camera learns simple pattern recognition. In a subsequent phase, the system is reset, the camera is rotated by a certain angle and the system learns from the same sample. An indifferent device should give identical results in both phases.*

In this experiment, the same system is used twice for the same learning task. The only difference is that the camera does not have the same orientation. We expect from an indifferent device that the outcomes are indistinguishable : after learning, the same images are assigned the same classes in both cases. Conversely, any difference in the learned classification would mean that the system is sensitive to some absolute orientation, *e.g.* that the system "knows" that the camera is vertical or horizontal. Such a system would be non-indifferent (in this case, non-isotropic).

Let us give a formal definition of indifference. A learning device as shown in figure 1 associates samples with binary classifications. If J is the training sample, $\mathcal{A}(J)$ is the learned classification. Each classification operates a partition of the N-hypercube into two classes. The learning device $\mathcal{A}$ is thus an application from the set of samples into the set of binary partitions of the N-hypercube.

A learning system $\mathcal{A}$, as defined figure 1, is *indifferent* w.r.t. isometries if its global behavior, including both learning and working phases, remains identical when inputs (*i.e.* examples *and* data) are systematically transformed through an isometry. In other words, if ρ is an isometry (*e.g.* a rotation) and x a test input, then $\mathcal{A}(\rho(J))$ puts $\rho(x)$ into the same class as $\mathcal{A}(J)$ does for x ($\rho(J)$ results from the application of ρ to every element of J). More precisely, a learning device $\mathcal{A}$ is *indifferent* if, for any isometry ρ from the N-hypercube into itself and for any sample J :

$$\mathcal{A}(J) = [\mathcal{A} \circ \rho(J)] \circ \rho$$

It can be shown that any isometry with respect to the Hamming distance in the N-hypercube results from the composition of a translation (which complements some given components of the binary vectors) and of a rotation (the effect of which is to permute components). Systems that are indifferent w.r.t. isometries are both isotropic and relative. *Isotropic* systems are indifferent w.r.t. any permutation of coordinates, and *relative* systems are indifferent w.r.t. any (partial) complementation. Since there are 2^N possible translations and $N!$ different permutations, the total number of isometries for the N-hypercube is $2^N \cdot N!$. It is possible to define the indifference property with respect to any group of transformations. The following results will still hold. The group of isometries seems however to be the most relevant in many cases. In particular, systems that are only sensitive to distances between data are indifferent to isometries.

4. Examples of indifferent mechanisms

Most statistical learning algorithms, including connectionist networks and Similarity Based Learning (SBL) algorithms, are indifferent or quasi-indifferent. For example, a Kohonen network [Kohonen 1984], completed with a decision device, is an indifferent system : both wiring and algorithm make reference to relative properties of the input only. Usual multi-layer perceptrons are almost indifferent : if one such perceptron, used in pattern recognition, is presented with “shuffled” input, *i.e.* if we systematically permute pixels of images to be learned and classified so as to make them unrecognizable to the human eye, the system behaves as it would have done with intact input[2]. The same happens if some pixels are inverted, *i.e.* if part of the input images is systematically negative. The only exceptions are permutations between data inputs and supervision inputs : after such a permutation, the system would probably be unable to learn the intended classification.

Most systems which, like the previous ones, learn by extracting statistical regularities from input are indifferent with respect to isometries. This holds, for instance, for usual Similarity Based Learning algorithms. Extraction of regularities is not a necessary feature of indifferent systems, however. Imagine a system that

[2] In the case of a Kohonen network, one should not confuse shuffled input with transformations in the Kohonen map. The former is represented by a mere point in the latter.

learns parity this way : it computes the parity to be learned by summing the *N* bits of one example (*K*=1), and then puts into class 1 all data having the same sum. Such a system is indifferent to any permutation (obvious) and to any complementation : if both example and data have some of their components systematically complemented, the result (class 1 or 0) will not change.

There are many different ways to be non indifferent. A crude example will be a device that computes an integer from the input (using the usual binary code) and decides class 1 iff the result is above the integer computed from the example (*K*=1). If the example is `0011` (here *N*=4) and the input is `0101`, the latter will be assigned class 1 since 5 is above 3. But after systematic permutation of first and third bit, we get `1001` for the example and `0101` for the data, and the decision will be class 0 since 5 is below 9. This system is not indifferent. There is an absolute reference that assigns *a priori* different roles to input components. Any system that makes use of *a priori* knowledge, as for instance structured matching systems [Ganascia 1987], has little chance to be indifferent.

We can see through these examples that indifference is not a quantitative measure of bias. For instance, the system learning parity from one example is indifferent, but is nevertheless strongly biased. It has an innate knowledge of parity. It has however no *a priori* knowledge of *even* and *odd*. What indifference checks is the absence of absolute bias.

5. The anisotropy-harmony dilemma

It can be shown [Dessalles 1993] that if an indifferent learning system can reach only a limited number of different classifications, then these classifications are necessarily harmonious. The *harmony* of a classification is the number of isometries which leave classes globally invariant. More precisely, for any indifferent system $\mathcal{A}$:

$$Harm(\mathcal{A}(J)) \cdot Var(\mathcal{A}(J)) = 2^{N}.N!$$

The harmony $Harm(\mathcal{A}(J))$ of a classification $\mathcal{A}(J)$ that has been learned from the sample J is inversely proportional to its *variety* $Var(\mathcal{A}(J))$, which is defined as $\mathcal{A}(\rho(J))\}$ where ρ is any isometry in turn[3]. One consequence of this result is :

If an indifferent learning mechanism $\mathcal{A}$ can reach only a limited number (<< 2^{N}x$N!$) of classifications, then these classifications are harmonious.

[3] The core of the demonstration lies in the fact that when a classification $\mathcal{C}$ can be reached by the learning system, $\mathcal{C} = \mathcal{A}(J)$, then all the other classifications obtained by isometric transformation from the classes of $\mathcal{C}$ are accessible as well, and can be written $\mathcal{A}(\rho(J))$.

This comes from the fact that if the system is able to learn only a few classifications, then *Var(𝒜(J))* must be small, and *Harm(𝒜(J))* must be high. We call it the anisotropy-harmony dilemma : when few forms can be learned, as in convergent learning, then either the learner is non-indifferent (anisotropic or non-relative) or the accessible forms are harmonious. This result has interesting consequences that we explore now.

6. Constraints on convergent learning

In many situations, we observe that different algorithms or organisms learn roughly the same things under various circumstances. Children learn to ride a bicycle the same way : they learn to turn the handlebars to the falling side by an appropriate angle, without trying to change pressure on pedals. They learn to speak their mother language in a way which is hardly distinguishable from the way of speaking of other children of the same school (pronounced phonemes, accent, syntactic forms used, etc.). They acquire roughly the same knowledge on a given subject (*e.g.* highway code). When very young, all of them draw trees perpendicular to the slope, and later draw them correctly [Piaget & Inhelder 1947:444]. Technical systems also learn reliably : different clustering algorithms (*e.g.* a moving center algorithm with different choices for the seeds) may come upon the same partition of a given set of data. A SBL algorithm may give the same characterization of classes when different (but coherent) sets of examples are given as input.

Convergence is an especially crucial requirement when the problem is to learn how to communicate. Learning introduces variety in communication. Contrary to bees, which are unable to vary their way of expressing the location of a food source, we *learn* how to express such things and many others. However, this only works because our fellows learned exactly the same code, and not a different one. In most situations relevant to social or evolutionary adaptability, we observe that there is “something” to be learned, and that learning individuals or species come most often upon similar solutions.

Now the question is, where does convergence come from ? The answer is quite obvious : either from exposure to quasi-identical data, or because a very small number of final sates are reachable. The first case is illustrated by biological convergence. For instance, dolphins and sharks evolved similar caudal fins. This convergence between two non related species results from physical constraints in the water medium that generates similar pressures on the body of both animals. Here, the identical “input” explains why two adaptive systems could converge towards similar states.

An opposite example is offered by Jean Piaget’s theory of learning. The core of this theory is the existence of a few definite states the child may reach [Piaget 1967]. These states are characterized by a set of operations the child can perform, which is closed for the combinations that the child is able to conceive. When this set reaches a group structure, then learning is complete (until a new kind of operation is discovered). For Piaget, all children go through the same states, and this is because there are only a few sets of actions which are closed for any combination.

How is convergent learning constrained by the anisotropy-harmony dilemma ? Whenever convergent learning is observed, then at least one of these alternatives must be true :

- there are many reachable forms
  - ➔ then convergence must follow from the reliability of data in the learning phase. Organisms must have been exposed to similar data (*i.e.* convergence is in the data).
- there are few reachable forms (*i.e.* convergence is a consequence of the organism's structure)
  - the learning mechanism is indifferent
    - ➔ then learned forms are necessarily harmonious.
  - the learning mechanism is not indifferent
    - ➔ then it possesses an *a priori* sensitivity to absolute features of input.

In other words, convergence is either in the data or is due to the learner's bias. In the latter case, the anisotropy-harmony dilemma applies : if what is learned is not harmonious, then an absolute innate reference must be postulated.

In cognitive modeling, when we are faced with a situation of convergent learning, we have first to check the reliability of data the organisms were exposed to. If there is no guarantee that the organisms had access to similar data, then we have to check the harmony of learned form. If it is low, then the learning mechanism is necessarily non indifferent. In other words, the organisms have some absolute bias towards a specific, non harmonious, form ; they are able to learn this form but not other, isometric, forms.

The presence of an absolute reference bias implies the existence of a specific innate component, but, as previously noticed, the converse is not true. The parity learning device imagined above is indifferent (*i.e.,* there is no absolute reference), but it has a strong innate "knowledge" of the way of computing parity. No wonder that parity is so easily learned by such an indifferent mechanism : both odd and even subsets of the hypercube are highly harmonious. The situation is much more interesting when some inharmonious form F_1 is reliably learned by a learning device under various circumstances. We are forced to conclude that this device has an absolute reference bias. It is so particular that it may be unable to learn a form F_2, isometric to F_1, whatever the kind of examples it is exposed to. For cognitive modeling purposes, it is thus interesting to check the harmony of learned forms, especially in the case of convergent human learning.

7. Human learning of harmonious forms

Human learning often seems to be convergent. To what extent are learned forms harmonious ? The idea that learned forms must be harmonious is at the root of the Gestalt theory. This theory insists on the importance of « good shapes », shapes that are simple, regular and symmetrical. For instance, the visual system is supposed to prefer the most regular and symmetrical perception which is compatible with sensory data. "Good" images, which can be described using less information, are recognized faster and are memorized better than odd ones [Rock & Palmer 1991]. Furthermore, any departure from symmetry is perceived as such and is analyzed as revealing the history of the object [Leyton 1992]. According to Gestalt theory, this holds not only for perception, but also for many abstract forms of learning, including the learning of conceptual knowledge, as Fritz Heider suggests [Rock & Palmer 1991]. This theory explains the convergence between mental processes acquired by different individuals by a preference for harmonious shapes. This leaves the door open for an indifferent learning mechanism.

We mentioned above that Piaget's theory predicts that only few forms can be learned. It never appeals to any reliability of data : all children have to perform operations in order to learn, but not necessarily exactly the same ones. The learning mechanisms invoked by Piaget (operational closure and the so-called "abstraction réfléchissante") appear to be strictly indifferent. We expect that accessible forms can be shown to be harmonious. Let us consider the well-known experiment of the two glasses. When water is poured from the wide glass into the narrow glass, children under six declare that there is now more water [Piaget 1967:610]. The young child does not take the section of the glass into account. For her, a correct situation is a situation which is compatible with what she knows of the effects of pouring (when pouring from the tap, the more water, the higher the level). This child would actually be surprised if pouring water into the glass caused lowering of the water level ! The set of « correct » situations can be described as $\{(V,h) \mid V=C \cdot h\}$, where V is the volume in the glass, h the height of water and C a constant. The set of situations that look correct for the older child, who considers the effect of the section of the glass, can be ideally described by $\{(V,h,\mathrm{r}) \mid \mathrm{V}=C' \cdot h \cdot r^2\}$, where r is the radius of the glass. This child would be amazed at any gross departure from this set. This experiment shows that each child learned a different form, the set of situations she considers as admissible. What is the harmony of such sets ?

These two sets are invariant for the operations each child is able to observe : $(V,h) \rightarrow (\alpha V, \alpha h)$ and $(V,h,r) \rightarrow (V, \alpha h, r/\sqrt{\alpha})$ respectively. These operations correspond to groups of isometries (translations) in logarithmic coordinates. The learned forms are thus highly harmonious. Piaget's interpretation of this experiment can be understood as the child making an extrapolation to the smallest harmonious (*i.e.* invariant for the accessible relevant operations) form.

Other learning mechanisms have been invoked to explain human learning capabilities. Many of them are by essence statistical and perform regularity extraction, especially those which are at the core of behaviorism : trials and errors, associationism, conditioning, etc. Such mechanisms are indifferent (or quasi-indifferent if they are supervised). When they are used under such situations that they generalize from a limited sample, what these mechanisms learn is quite harmonious : for instance the learned sets are invariant for all transformations affecting components which vary among examples. Otherwise, when such systems do not generalize, in other words if overfitting occurs, learned forms may vary considerably. We cannot say anything about the harmony of these forms. Convergence, if any, must then result from the reliability of data. In the case of language acquisition, this constraint has often been acknowledged (*e.g.* [Plunkett & Marchman 1990]) or presented as problematic by some authors [Piatelli-Palmarini 1988] who mention the “poverty of the stimulus”.

8. Human learning of inharmonious forms

Some aspects of visual perception are claimed to be indifferent. Stratton’s well-known experiments show that if you see the world upside-down through special glasses, then after a week without removing the glasses the world no longer looks weird [Gregory 1966]. This strongly suggests that visual perception, when acquired by newborns, may be indifferent to 180° rotation. However, we cannot conclude that our visual system is not *a priori* sensitive to absolute parameters. Would children see the world as normal if images presented to them were negative, or shuffled ? Such transformations are isometries in the visual space, because they do not change distances between images (*e.g.* similar images have similar negatives). Psychologists showed that our perception relies on absolute preferences. For instance, we are sensitive to figure-ground contrast [Wertheimer 1923]. Psychologists also gave evidence showing children’s preferences for whole objects when learning word meaning [Markman 1990]. Figure-ground contrast and object continuity disappear when images are shuffled, and we may predict that children living with shuffled perception would not be able to recover them. Another illustration of absolute bias in visual perception is given by the specific processing devoted to face recognition. A small region of our brain, located in the parieto-occipito-temporal area, seems necessary for the recognition of familiar faces [Tranel & Damasio 1985]. Again, we hardly imagine that such a system could operate on shuffled images and that children could develop normally in a world peopled by Picasso-like faces.

Some mechanisms that have been suggested to explain convergent human learning of language depart explicitly from indifference. In the modular theory of cognition advocated by Fodor [1983], many basic cognitive functions are performed by dedicated *modules* with an innate component that puts strong constraints on what can be learned by the child. Face recognition could be one such module, but the archetypal example is certainly language processing. N. Chomsky [1968, 1988] asserts that humans are innately biased to learn a small number of language structures. The corresponding learning mechanism, the setting of parameters

[Lightfoot 1991, Crain 1991] is by no means indifferent : it relies on a matching with preexisting structures. Imagine that in a remote area, natives speak a strange language : it is like English, except that some words are systematically permuted in each sentence, for instance the first and the fourth. Correct sentences in this “language” would be :

dark girl with the hair holds the baby rabbit

cars were many there in the garage

and did Mary what you do on the way home ?

No linguist would accept such word strings as examples of any possible human language, even if their meaning may be somehow recovered. They violate a basic principle that prevents syntactic constituents from partially overlapping. In the first example, the prepositional “phrase” [*dark _ with _ air*] and the noun “phrase” [*girl _ the*] overlap. Such a transformation that preserves surface similarity between word strings would dramatically affect children’s ability to learn their mother language. This learning process is not indifferent. As a consequence, we do not expect the set of grammatical sentences to show symmetry, *i.e.* invariance through surface transformations. Surface similarity is of little help to determine which sentences are syntactically correct [Piatelli-Palmarini 1988]. The same is true for meaningful sentences, at the semantic level [Jackendoff 1983] or, at the pragmatic level, for relevant utterances in a given situation [Dessalles 1993]. To account for language acquisition by an indifferent mechanism, one has to invoke reliable data. This solution has been resolutely criticized by Chomskyans who insist on “the poverty of the stimulus”, *i.e.* the lack of reliability of the input children are exposed to [Piatelli-Palmarini 1988, Pinker 1994]. The learning mechanism put forward by Chomsky (matching with innate structures and parameter setting), being highly anisotropic, does not require human languages to be harmonious, despite the alleged relative small number of target structures.

9. Conclusion

We have proposed here an original way of characterizing learning mechanisms, based on geometrical considerations. We defined the property of indifference, which captures a common feature of many usual learning models that are both isotropic and relative. A consequence of this definition is the anisotropy-harmony dilemma : when input is not rigidly invariable, isotropic or indifferent systems learn only harmonious forms reliably.

The consequences of this dilemma can be observed in cognitive modeling. Indifferent models of learning, like connectionism, behaviorism, Piaget’s theory, Gestalt theory, etc. predict harmonious results. According to such models, learning proceeds through a generalization towards the closest harmonious form compatible with input data. Such models may be an accurate account of many human learning abilities. However, they are not good candidates to explain reliable learning of inharmonious forms, or when the learning ability to be modeled is suspected to be non-isotropic. Aspects of visual pattern recognition (*e.g.* figure-ground contrast, whole object assumption, face recognition) and language are good examples of cognitive abilities that cannot be explained by indifferent mechanisms : the learning process is sensitive to many isometric transformations of input, and the target forms are not harmonious.

By acknowledging the importance of the property of indifference, one can think of a new approach to cognitive modeling : by systematically checking all the transformations leaving the learning process indifferent, one will get constraints on what the learning mechanism can or cannot be. If a model (*e.g.* connectionism) allows for greater indifference than observed, then it must be modified or ruled out, even if it reproduces the learner's performance accurately.

Isotropic and relative bias is generally preferred to avoid unnecessary specificity. However, indifferent learning mechanisms are specialized in some way : when data are not strictly reliable, these systems are bound to learn harmonious forms. The property of indifference appears to be a significant parameter that should be systematically taken into account in cognitive modeling.

Acknowledgments : I thank Eric Bonabeau, Gérard Cohen, Olivier Hudry and François Yvon for their help.

References

Boucheron, S. (1992). *Théorie de l'apprentissage*. Paris : Hermès.

Chomsky, N. (1968). *Le langage et la pensée (Language and mind)*. Paris : Payot, ed. 1969.

Chomsky, N. (1988). *Language and problems of knowledge*. Cambridge : The MIT Press, ed. 1992.

Crain, S. (1991). "Language Acquisition in the Absence of Experience". *Behavioral and Brain Sciences*, *14*, 597-650.

Dessalles, J-L. (1993). *Modèle cognitif de la communication spontanée, appliqué à l'apprentissage des concepts - Thèse de doctorat*. Paris : ENST - 93E022.

Elman, J. L., Bates, & et al. (1997). *Rethinking innateness*. Cambridge : M.I.T. Press.

Fodor, J. A. (1983). *La modularité de l'esprit*. Paris : ed. de Minuit, ed. 1986.

Ganascia, J-G. (1987). "AGAPE: De l'appariement structurel à l'apprentissage". *Intellectica*, *2/3*, 6-27.

Gregory, R. L. (1966). *Eye and brain - The psychology of seeing*. London : Weidenfeld & Nicolson, ed. 1977.

Jackendoff, R. (1983). *Semantics and Cognition*. Cambridge : The MIT Press, ed. 1995.

Kohonen, T. (1984). *Self-Organization and Associative Memory*. Berlin : Springer Verlag, ed. 1988.

Leyton, M. (1992). *Symmetry, causality, mind*. Cambridge MA : The MIT Press.

Lightfoot, D. (1991). *How to set parameters*. MIT Press.

Markman, E. M. (1990). "Constraints Children Place on Word Meanings". *Cognitive Science*, *14*, 57-77.

Piaget, J. & Inhelder, B. (1947). *La représentation de l'espace chez l'enfant*. Paris : P.U.F., ed. 1972.

Piaget, J. (1967). *Biologie et connaissance*. Paris : Gallimard.

Piatelli-Palmarini, M. (1988). "Evolution, selection and cognition: From 'learning' to parameter setting in biology and in the study of language". *Cognition*, *31*, 1-44.

Pinker, S. (1994). *The language instinct*. New York : Harper Perennial, ed. 1995.

Plunkett, K. & Marchman, V. (1990). *From Rote Learning to System Building*. San Diego : CRL Technical Report 9020, Univ. of California.

Rock, I. & Palmer, S. (1991). "L'héritage du gestaltisme". *Pour La Science*, *160*.

Schaffer, C. (1994). "A conservation law for generalization performance". In (ed), *Proceedings of the Machine Learning 1994 Conference*. Rutgers University, 259-265.

Tranel, D. & Damasio, A.R. (). "Knowledge without awareness: an automatic index of facial recognition by prosopagnosics". *Science*, *228*, 1453-1454.

Wertheimer, M. (1938). "Laws of organization in perceptual forms". In Willis D. Ellis (ed), *A source book of Gestalt psychology*. London : Routledge & Kegan, 71-88.

Limites des modèles isotropes de l'apprentissage naturel et artificiel

Jean-Louis Dessalles [1]

Tous les systèmes d'apprentissage présentent un biais. Le système parfait et universel qui éviterait toute prédétermination « innée » n'existe pas. Pour autant, tous les biais ne sont pas équivalents. On préfère, en général, éviter les biais *anisotropes* lorsque l'on conçoit un mécanisme d'apprentissage, surtout lorsqu'il est supposé modéliser une compétence cognitive humaine ou animale. Un biais anisotrope suppose nécessairement quelque connaissance *ad hoc* qui serait possédée *a priori* , ce qui limite la généralité du système apprenant.

Nous voulons suggérer toutefois que les modèles d'apprentissage *isotropes*, malgré leur apparente généralité, peuvent se révéler trop limités. Dans de nombreux cas, les individus d'une même espèce vivante apprennent des formes identiques dans des circonstances variées. Nous montrons que ces situations d'apprentissage, dites convergentes, ne peuvent pas être expliquées par des modèles isotropes, à moins que les formes apprises soient harmonieuses, c'est-à-dire qu'elles aient de fortes propriétés de symétrie. Ce *dilemme anisotropie-harmonie* dérive d'une caractérisation formelle du biais, basée sur des propriétés géométriques simples. En montrant comment ce dilemme affecte les théories classiques de l'apprentissage, nous tentons de clarifier le débat classique entre l'inné et l'acquis dans le cas de l'apprentissage convergent.

mots clés : Apprentissage, isotropie, biais, symétrie, inné.

[1] ENST - Département Informatique - 46 rue Barrault - 75013 PARIS - France - courriel : dessalles@enst.fr