In D. Canamero & M. van Someren (ed), *Learning in humans and machines - ECML'98 Workshop notes*. Chemnitz : Technische Universität Chemnitz - CSR-98-03, 6-17. (1998)

Characterising Innateness in Artificial and Natural Learning

Jean-Louis Dessalles

Ecole Nationale Supérieure des Télécommunications
46 rue Barrault - F-75013 Paris - France
E-mail : dessalles@enst.fr

Abstract. The purpose of this paper is to propose a refinement of the notion of innateness. If we merely identify innateness with bias, then we obtain a poor characterisation of this notion, since any learning device relies on a bias that makes it choose a given hypothesis instead of another. We show that our intuition of innateness is better captured by a characteristic of bias, related to isotropy. Generalist models of learning are shown to rely on an "isotropic" bias, whereas the bias of specialised models, which include some specific *a priori* knowledge about what is to be learned, is necessarily "anisotropic". The so-called generalist models, however, turn out to be specialised in some way: they learn "symmetrical" forms preferentially, and have strictly no deficiencies in their learning ability. Because some learning beings do not always show these two properties, such generalist models may be sometimes ruled out as bad candidates for cognitive modelling.

1 Introduction

In cognitive modelling, it is generally considered more parsimonious to avoid *a priori* knowledge of aspects of what is to be learned. 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. In several cases, a specialised connectionist architecture can reproduce the child's performance. The design of such architectures is suggested by general principles, it does not include an *a priori* knowledge of what is to be learned, contrary to what is claimed in other theories of learning (Chomsky, 1968). This does not mean that the connectionist devices used by these authors are not "tuned" for the task they perform. Each relies indeed on a bias that allows it to correctly induce. But this bias seems more legitimate than a direct knowledge of the target, which would appear as a kind of "cheating". We will propose a notion, which we call *indifference*, that captures this intuition of what is legitimate bias and what is not.

The work presented here is based on qualitative, geometrical, considerations and does not refer to cost functions, sample probability or asymptotic correctness. Learning Theory is most often concerned with the problem of making good inductions, of finding a way of choosing a good hypothesis among a given set of classifications. When performing cognitive modelling, however, one may adopt a different approach. Our problem here is not to investigate how learning systems may approximate a given regularity. We will rather look for a link between some intrinsic

properties of the learning device (*esp.* isotropy) and properties of the classifications it may learn. From this link we will infer a characterisation of the device's innate knowledge.

First we will define the property of *indifference*, a notion that includes isotropy. Indifference appears to be a very common feature of learning devices. We also define *harmony* as a property which is associated with the symmetry of classifications to be learned. Then we establish a connection between these apparently independent notions: under certain circumstances, indifferent mechanisms are bound to learn harmonious classifications. We will discuss the consequences of this result by briefly reviewing some important cognitive learning mechanisms (Gestalt Theory, Piaget's Theory, Associationism, Inneism) to show how they comply with this constraint.

2 Isotropy and Indifference of Learning Mechanisms

2.1 A Simple Learning Device

In order to illustrate the notions defined here, we will consider a generic learning device that learns binary classifications (Figure 1). Most learning devices can be analysed in a similar way.

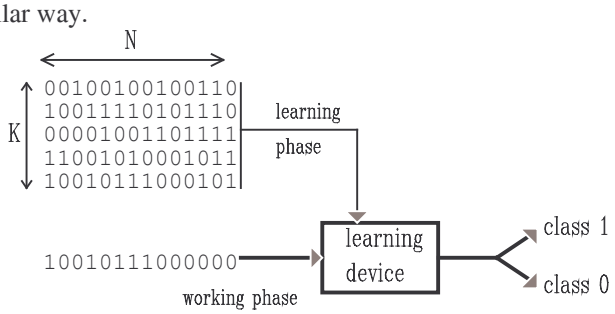

Fig. 1. Generic learning mechanism

Input is given through N binary sensors. During the learning phase, the system is presented with a sample of K examples of N bits (some components of these examples may represent supervision information if any). During the subsequent working phase, the system is able to assign one of two classes (class 1 or class 0) to any binary N-uple. Each learned binary 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.

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. For our purpose here, there is no need for assessing the adequacy of the learned classification to pre-existing classes (contrary to PAC learning studies). Moreover, such a measurement may be problematic in certain situations encountered in cognitive modelling. 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. It is nevertheless possible, without considering accuracy, to tell sometimes what a learning system cannot be. We will suggest that some cognitive learning performances are not the result of isotropic, or indifferent mechanisms.

2.2 Indifference as Absence of Absolute Reference

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, which works as an absolute reference, is in some way "hard-wired" in the system. 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.

We can give an intuitive idea of the indifference property by considering a simple pattern learning experiment. Imagine that a device using a camera can be trained to recognise simple patterns like, say, written digits. In the learning phase, the system is presented with examples of written shapes. In the working phase, the system classifies all patterns it perceives in two classes, digits and non-digits. Whatever the errors it makes, we are interested in the classification $\mathcal{C}$ that has been learned by this system. Now imagine that the memory of the system is reset, and that the camera is rotated by a given angle. We present it with the same examples as previously, in the same order. Then we check what classification it has learned, keeping of course the camera in its new position. If the system is indifferent, then the result is $\mathcal{C}$.

The learning system $\mathcal{A}$, as defined figure 1, is *indifferent* w.r.t. isometries if its global behaviour, including both learning and working phases, remains identical when inputs (*i.e.* examples *and* data) are systematically transformed through an isometry. 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:

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

where $\rho(J) = \{\rho(x) \mid x \in J\}$. The symbol $\circ$ stands here for the composition of functions. In the terms of our previous experiment, $\rho(J)$ is the set of examples seen through the rotated camera. $\mathcal{A}(\rho(J))$ is the corresponding learned classification. When this classification is presented with rotated data (data seen through the rotated camera), the class assigned is the same as in the initial experiment, in which the camera was not rotated. In other words, if x is a written pattern, $\rho(x)$ is given the same class by $\mathcal{A}(\rho(J))$ as x was given by $\mathcal{A}(J)$.

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). We can thus define an *isotropic* system as being indifferent to

any permutation of coordinates. A system will be said to be *relative* if it is indifferent to any (partial) complementation. Since there are 2^N possible translations and $N!$ different permutations, the total number of isometries for the N-hypercube is $2^N \times N!$. 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.

2.3 Examples of Indifferent Learning Devices

The notion of indifference is based on purely geometrical considerations, namely the insensitivity to isometries. It does not depend on the choice of a learning accuracy measure. As a consequence, it can be used to characterise a large variety of learning situations, as we will see.

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 multilayer perceptrons are almost indifferent: they are indifferent to all isometries, except for permutations between data inputs and supervision inputs. Imagine a system that 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. More generally, most connectionist systems and most statistical learning devices will be indifferent with respect to isometries. This holds for instance for usual Similarity Based Learning algorithms.

There are many different ways to be non indifferent. A crude example would 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 the size of input is N = 4 bits) and the datum is `0101`, the latter will be assigned class 1 since 5 is above 3. But after permutation of first and third bit from the left, we get `1001` for the example and `0101` for the datum, 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 coordinates. Any system that makes use of *a priori* knowledge, as for instance structured matching systems (Ganascia, 1987) has little chance to be indifferent.

3 Limits of "Generalist" Learning models

Learning systems which are claimed to be generalist, *i.e.* which are supposed to learn under a large variety of circumstances, are generally indifferent. This is true of most statistical algorithms, connectionist systems included. A non-indifferent learning ability would prevent the learner from behaving similarly under equivalent (esp. isometric) situations. Indifferent systems, which do not show this limitation, show however another kind of restriction. We propose to show that when an indifferent learning system can reach only a limited number of different classifications, then these classifications are necessarily harmonious.

Harmony is a more general notion than symmetry. The *harmony* of a subset of the *N*-hypercube is the number of isometries (or more generally of transformations among those considered) which leave the subset globally invariant. The harmony of a classification is the harmony of the two classes it defines. The following result holds for any indifferent system $\mathcal{A}$:

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

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 the cardinal of $\{\mathcal{A}(\rho(J))\}$ where ρ is any isometry (or any considered transformation) in turn. Indications about the proof are given in annex.

One important consequence of this result is:

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

This comes from the fact that if few classifications can be learned, $Var(\mathcal{A}(J))$ is small, smaller than the total number of learnable forms. As a consequence, $Harm(\mathcal{A}(J))$ is high.

This result seems to undermine the notion of "generalist" learner. Such systems should be indifferent, but as such they must comply with the above constraint. As we will see, it proves to be a quite strong limitation.

The absence of really generalist learners is predicted by Schaffer's conservation law, which states that a learning device $\mathcal{A}$ cannot be always good: its generalisation accuracy is equivalent to random when averaged on all possible pre-existing regularities (Schaffer, 1994). In other words, any learner must have its preferred learning situations in which it performs efficiently. Our result shows that indifferent systems perform well on harmonious forms. This result has important consequences for the study of innateness, that we will explore now.

4 Two Forms of Innateness

We can now tell the difference between two forms of innateness: indifferent and non-indifferent bias. In what we call *convergent* learning, which is certainly the most interesting situation to be observed, both biases show different behaviours. This is particularly relevant to cognitive modelling, since we may infer the type of bias from the learning performance.

4.1 Convergent Learning

We often observe that different algorithms or organisms reliably learn roughly the same forms under various circumstances. For instance, different clustering algorithms (*e.g.* a moving centre algorithm with different choices for the seeds) may come upon the same partition of a given set of data. A Similarity Based Learning algorithm may give the same characterisation of classes when different (but coherent) sets of examples are given as input. 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 the pressure on pedals. They learn to speak their mother language in a way which is hardly distinguishable from the way other children of the same school speak (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).

Convergence is an especially crucial requirement when the problem is to learn how to communicate. For sure learning allows variety in communication. Bees are unable to vary their way of expressing the location of a food source. Contrary to them, we *learn* how to express such things and many others, but this only works because our fellows learned exactly the same code, and not a different one.

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. For instance in the case of the bicycle, we would opt for the second hypothesis. The core of Jean Piaget's theory is the existence of definite *stable* states the child may reach (Piaget, 1967) . These states are characterised 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.

4.2 Convergence and Innateness

As we can infer from the preceding results, when convergence of learning results from the small number of reachable forms, then these forms are inexorably *harmonious* if the learning mechanism is indifferent. We are thus faced to the following dilemma. If there is convergence in a learning process, then at least one of these alternatives must be true:

- there are many reachable forms
  then no specific innate component can be inferred. Convergence results from the reliability of data in the learning phase (*i.e.* convergence is in the data).
- there are few reachable forms (*i.e.* convergence is a consequence of the organisms' structure)
  if these forms are inharmonious
  then the learning mechanism *cannot* be indifferent: it possesses an innate sensitivity to absolute features of what is to be learned.
  if these forms are harmonious
  then an indifferent learning mechanism should be suspected

In cognitive modelling, when we are faced to 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. Such organisms have an 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 (section 2.3) is indifferent (*i.e.,* there is no absolute reference), but it has a strong innate "knowledge" of the way of computing parity. It is not a 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 device under various circumstances. We are forced to conclude that such a 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 modelling purposes, it is thus interesting to check the harmony of learned forms. Especially in the case of convergent human learning.

5 The Question of Innateness in Convergent Human Learning

5.1 Learning Harmonious Forms

The preceding development allows us to ask the question of innateness differently: is the innate component that necessarily underlies human learning performance "indifferent" or not ? First, we must check to what extent learned forms are harmonious.

The idea that learnable 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 faster recognised and are better memorised than odd ones (Rock & Palmer, 1991). Furthermore, any departure from symmetry is perceived as such and is analysed as revealing the history of the object (Leyton, 1993). According to the 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 it (Rock & Palmer, 1991). This theory explains the convergence between mental processes acquired by different individuals by the existence of these harmonious shapes. This strongly suggests that an indifferent learning mechanism is involved.

We mentioned above that Piaget's theory predicts a few learnable forms. It never appeals to 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 as strictly indifferent. No wonder therefore 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 $\{(V,h) \mid V=C\times h\}$, 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 described by $\{(V,h,r) \mid V=C' \times hr^2\}$, 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 (regularity extraction, 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 *generalise* from a limited sample, learned forms are quite harmonious (for instance these forms are invariant for all transformations affecting components which vary among examples). When there is no generalisation, *i.e.* if overfitting occurs, learned forms may vary considerably. We cannot say anything about the harmony of these forms, but convergence must result in this case 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, as we will see.

5.2 Learning 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 new-borns, 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. 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

with beautiful girl the dark hair holds the baby rabbit

girl did the when with dark hair hold the baby rabbit?

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 this way. 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). To account for language acquisition by an indifferent mechanism, one would have to invoke reliable data. This solution has been resolutely criticised 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.

6 Conclusion

We have proposed here an original way of characterising innateness in learning mechanisms, based on geometrical considerations, and independently of any accuracy measure. We defined the property of indifference, which captures a common feature of many usual learning models that are isotropic and relative. Then we suggested that indifferent learning mechanisms were constrained. Roughly speaking, an indifferent mechanism cannot lead to convergent learning if neither the forms that are learned are harmonious (*i.e.* invariant for many isometries) nor data are reliable (*i.e.* similar for different learners).

The consequences of this result can be observed in cognitive modelling. Indifferent models of learning, like connectionism, behaviorism, Piaget's theory, Gestalt theory, etc. postulate no specific *a priori* knowledge of the forms that are eventually acquired. Correspondingly, they predict harmonious results. According to such models, learning proceeds through a generalisation 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 modelled 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 modelling: 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 specialised 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 modelling.

Acknowledgements: I thank Eric Bonabeau, Gérard Cohen and Olivier Hudry for their help.

References

Chomsky, N. (1968). *Language and mind*. French ed. Paris: Payot, 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.

Elman, J. L., Bates, E. A., Johnson, M.H., Karmiloff-Smith, A., Parisi, D. & Plunkett, K. (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.

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. & et al., (1967). *Logique et connaissance scientifique*. Paris: Gallimard (La Pléiade), ed. 1976.

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". *Proceedings of the Machine Learning 1994 Conference*. Rutgers University, 259-265.

Tranel, D. & Damasio, A.R. (1985). "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 W. D. Ellis (ed), *A source book of Gestalt psychology*. London: Routledge & Kegan, 71-88.

Annex

For any indifferent system $\mathcal{A}$:

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

The *harmony* of a subset of the *N*-hypercube is the number of isometries which leave the subset globally invariant. The harmony of a classification is the harmony of the two classes it defines. The core of the demonstration lies in the fact that when a classification $\mathcal{C}$ can be reached by the leaning 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 $\mathcal{A}(\rho(J))$. Here are the main steps of the proof:

- In order to see what happens to $\mathcal{A}(\rho(J))$ when ρ is any isometry, we define the following equivalence relation $\leftrightarrow$ in the set of reachable classifications:

$$\mathcal{A}(J_1) \leftrightarrow \mathcal{A}(J_2) \quad \text{iff } \exists\, \rho \text{ isometry}, \quad J_2 = \rho(J_1) = \{\rho(x) \mid x \in J_1\}$$

- If $\mathcal{A}$ is an indifferent mechanism, we can see that ρ acts as a bijection between class 1 (resp. class 0) defined by $\mathcal{A}(J_1)$ and class 1 (resp. class 0) of $\mathcal{A}(J_2)$.
- If an isometry σ leaves class n°i of $\mathcal{A}(J_1)$ invariant, then the isometry $\rho \circ \sigma \circ \rho^{-1}$ leaves class n°i of $\mathcal{A}(J_2)$ invariant. As a consequence, $Harm(\mathcal{A}(J_1)) = Harm(\mathcal{A}(J_2))$.
- By definition of the variety, $Var(\mathcal{A}(J_o))$ is equal to $cardinal(C(J_o))$ where $C(J_o)$ is the equivalence class of $\mathcal{A}(J_o)$ for $\leftrightarrow$. We note $Inv(\mathcal{A}(J_o))$ the set of isometries leaving $\mathcal{A}(J_o)$ invariant. Any isometry either is in $Inv(\mathcal{A}(J_o))$, or changes $\mathcal{A}(J_o)$ into another classification equivalent for $\leftrightarrow$.
- For a given J_o, we define an equivalence relation $\mathbin{⌘}$ among isometries: $\rho \mathbin{⌘} \tau$ iff $\mathcal{A}(\rho(J_o)) = \mathcal{A}(\tau(J_o))$. It can be easily shown that each equivalence class of $\mathbin{⌘}$ can be written

$\rho \square o\ Inv(\mathcal{A}(J_o))$, with ρ being an appropriate isometry. In particular, all equivalence classes for $\mathfrak{R}$ have the same number of elements $cardinal(Inv(\mathcal{A}(J_o)))$.

- We can see that $\rho \square o\ Inv(\mathcal{A}(J_o)$, can be rewritten $Inv(\mathcal{A}(\rho(J_o))$. But $\mathcal{A}(\rho(J_o)) \leftrightarrow \mathcal{A}(J_o)$. Therefore each equivalence class for $\mathfrak{R}$ corresponds to a different element of $C(J_o)$. There are $cardinal(C(J_o))$ different equivalence classes for $\mathfrak{R}$, each of them having $cardinal(Inv(\mathcal{A}(J_o)))$ elements. Since there are globally $2^N \cdot N!$ different isometries, we get $cardinal(Inv(\mathcal{A}(J_o))) \cdot cardinal(C(J_o)) = 2^N \cdot N!$, which was the relation to be proven.