Models of abduction

Paul Bourgine
bourgine@poly.polytechnique.fr
CREA-Ecole Polytechnique

"*the mind's capacity to guess the hypothesis with which experience must be confronted, leaving aside the vast majority of possible hypotheses without examination*" (Peirce)

Abstract:

The aim of this paper is to sketch a theory of abduction with its relations with deduction and induction in the sense of the second Peirce. The relation of abduction is seen as a relation which is in some sense reciprocal to deduction. Furthermore, this conception of abduction is directly compatible with the conception of induction underlying belief revision.

The main result is a representation theorem which unifies the various meanings of *abduction*. This theorem has two faces, one axiomatic and the other geometric, which are equivalent. This meaning of abduction is expressed through models from the three main paradigms of cognition; these models aim at establishing its adequation for the cognitive science.

Introduction:

C.S. Peirce had the view that reasoning involves three kinds of inference, abduction, deduction and induction. His reflections on this question developed over forty years, at a time when logic in its modern guise started emerging. Peirce's thoughts changed substantially during this period, as was shown by Burks, Fann, Thagard and Anderson. These analysts distinguished two main phases in Peirce's reflections about the kinds of reasoning. In the first phase, before 1900, Peirce offers a syllogistic approach of the three kinds of reasonings; in the second phase, after 1900, he favors an inferential approach which is closer to scientific inquiry.

In the syllogistic approach, exemplified by the celebrated Barbara syllogism, deduction is the type of reasoning which allows deriving, from a major premiss (the rule) and a minor premiss (the case), a conclusion (the result). Induction derives by generalization a rule from a collection of observations of case-result pairs. Abduction starts with the conclusion and the major premiss to derive the minor premiss.

In the inferential approach, the function of abduction is to emit an hypothesis, which can be either a premiss, a rule or a theory. The function of deduction is to draw from an hypothesis its necessary or probable consequences. The function of induction consists in comparing the predicted consequences with the observed results. These three kinds of reasoning are considered necessary in scientific inquiry, abduction being the main instrument in a logic of discovery. When a new surprising fact is encountered, the first stage in reasoning consists in abducing an explanatory hypothesis, the second in deducing the testable consequences and the third in testing these consequences to either confirm or falsify the explanatory hypothesis. In its inferential approach, the second Peirce distinguishes clearly these three kinds of inferences and makes their relations explicit.

The aim of this paper is to sketch a theory of abduction with its relations with deduction and induction in the sense of the second Peirce. Abduction is seen as a relation reciprocal to deduction, not directly as done by Flach [Flach, 96], but in a more general and sophisticated sense very close to the framework of belief revision [Alchourron & al., 85; Gärdenfors, 88; Katsuno & al., 91]. One supplementary advantage beyond generality is that this conception of abduction is directly compatible with the conception of induction underlying belief revision.

This theory of abduction will be expressed within models belonging to different paradigms of cognition , the purpose being to check its adequacy for the whole field of cognitive science. Three main paradigms are considered : the cognitivist paradigm which takes knowledge to be symbolic, with validity as a criterium of success; the connectionist paradigm which substitutes subsymbolic states of a neural network to symbols, while retaining the same criterium of success ; the constructivist paradigm which replaces the criterium of validity with an evolutionary criterium of viability and claims that the symbolic level to be grounded in the subsymbolic level. The three parts of the paper are discussing models of abduction corresponding to these three paradigms. For the sake of clarity, inferences are not considered to be probabilistic and are only analyzed in a set-theoretic framework, which supposes that t Nature‘s responses are deterministic.

1. Abduction and cognitivism

Basic schema

Let us consider a special type of expert, a physician. There is a set of diseases which are not directly observable ; only signs are available to the physician. A causal relation develops from a hidden disease to a set of observable signs. The physician’s reasoning moves in the converse sense from the observable signs towards the hidden diseases : she has to « abduce » the hidden hypothesis (disease) from the observable facts (signs).

What is first to be understood is the constraints that apply to this kind of reasoning. The most convenient way to express these constraints, as is the case in other kinds of reasoning, is to spell out the axioms which abduction follows. A great advantage of such a specification of abductive reasoning is that it can be falsified by psychological experiments. More precisely, a set of axioms defines a class of abductive reasonings which can be enlarged by weakening the axioms or restricted by strengthening them. The above set of axioms seems to be both interesting for it global properties and plausible :

- A1-Consistency : nothing can be abduced from a contradiction. It is the dual of the well-known property of deduction : everything can be deduced from a contradiction.

- A2-Success : every hypothesis can be abduced from itself. Again, this is the dual of the property that every hypothesis can be deduced from itself.

- A3-Cautious monotony : if one can abduce a first hypothesis from a fact and if this hypothesis can be abduced from an other hypothesis, one can abduce both hypotheses from the fact. This principle can be reformulated by saying that the abductive inference for an hypothesis can be extended to other underlying facts or hypotheses which might confirm the hypothesis.

- A4-Or : if one can abduce a first hypothesis from a fact and if a second hypothesis is incompatible with the first one then the conjunction of the two hypotheses can be abduced from the conjunction of the fact and the second hypothesis. This principle can be reformulated by saying that if many incompatible hypotheses are possible, they can be preserved within a particular scheme of abduction.

- A5-And : If one can abduce a hypothesis from two facts then one can abduce it from the conjunction of these two facts.

We suppose that the facts and the hypotheses are represented as propositions constructed with the help of the classical connectors $\{\neg, \wedge, \vee, \rightarrow, \leftrightarrow\}$ from a finite (for the sake of simplicity) set of atomic propositions P={ $\pi 1, \ldots, \pi n$ }. The set W of possible worlds is the set of interpretations of P, i.e. the set of functions from P to {T=true, ^=false}. To each proposition « a » it is possible to associate the set of possible worlds «A » which makes the proposition true. Then the syntactic implication a $\rightarrow$ b holds if and only if the semantic inclusion A $\subseteq$ B holds between their corresponding sets of worlds «A » and «B ». And the syntactic equivalence a $\leftrightarrow$ b holds if and only if the semantic equality A=B holds.

In this framework, we now collect together the previous axioms for abductive inferences, where «a |< b » means « from a it possible to abduce b » or simpler « from a one abduces b » :

A1-Consistency : $\neg$(^ |< a)

A2-Success : a |< a

A3-Cautious monotony : if a |< b and g |< b then a |< b $\wedge$ g

A4- Weak Or : a |< b and b $\wedge$ g $\rightarrow$ ^ then a $\vee$ g |< b $\vee$ g

A5-And : a |< g and b |< g then a $\wedge$ b |< g

A deep representation theorem allows to give the canonical form of abductive inferences satisfying A1-A5. This representation theorem is given in two forms : theorem 1 links abduction and deduction ; theorem 1' prepares the link between abduction and induction, through the belief revision operator.

Theorem 1 (representation theorem for abductive reasoning)1:

The abductive inference « |< » follows the axioms A1-A5 if and only if there is a total preorder relation « $\leq$ » on the worlds set W such that

(i) a |< b iff a |< min(b) for any b

(ii) a |< b iff b |- a when b= min(b) (b is said « parsimonious »)

Given a preorder relation « $\leq$ » on the worlds set W, the meaning of min(b) is clear : it consists in selecting the set B of worlds where b is true, keeping its minimal worlds min(B) ={w $\in$ B :w' $\leq$ w for all w' $\in$ B}and returning to the corresponding proposition min(b). In this theorem, the concept of a parsimonious hypothesis is crucial : it is simply an hypothesis which contains only its minimal worlds. Now, point (i) means that the only relevant hypotheses in abductive reasoning are the parsimonious hypotheses, and point (ii) means that abduction is the relation reciprocal to deduction for parsimonious hypotheses.

It is easy to verify that an inference relation which satisfies points (i) and (ii) when there is a preorder relation in the set of possible worlds is an abductive inference in the meaning of axioms A1-A5. What the theorem of representation expresses is that all the abductive inference satisfying A1-A5 have necessarily this canonical form.

The total preorder « $\leq$ » defines a system of spheres on the finite set W of possible worlds. The inner sphere is the set K of minimal worlds. Then comes the sphere K1 of minimal worlds of W-K, and K2 of the minimal worlds of W-K-K2,… These spheres are also defined as the equivalent classes of W by the equivalent relation « $\approx$ » defined by the preorder (w$\approx$w' iff w$\leq$w' and w'$\leq$w). Thus it is possible to define a distance from a set to the set K: the worlds in K1 are at distance « 1 » from K, in K2 at « 2 », ... And min(B), which is the set of minimal worlds of B, is also the set of worlds of B with minimal distance from K, which is written as K*B in the belief revision literature. And a |< b iff K*B $\subseteq$ A.

Theorem 1' (representation theorem for abductive reasoning):

The abductive reasoning « |< » follows the axioms A1-A5 if and only if there is a total preorder relation « $\leq$ » on the worlds set W with minimal worlds K such that a |< b iff K*B $\subseteq$ A where K*B is the set of the nearest from K worlds of B : K*B ={w$\in$B : d(w,K)=d(B,K)}.

These representation theorems have a deep meaning by linking in both directions an axiomatic sense and a geometrical sense for abduction : if the abductive reasoning « |< » follows the axioms A1-A5 (which can be falsified), then the beliefs have necessarily the above topology of spheres, with the preferred beliefs in K and then less and less preferred beliefs in the farther and farther spheres. Furthermore, additional properties of abduction follow from the previous axioms. Some of the most important are listed in the following proposition :

Proposition 2 : if the abductive reasoning « |< » follows the axioms A1-A5 then :

A1'-Falsification : if a |< b and |=a $\wedge$bg then $\neg$(a $\wedge$ $\neg$g |< b)

A2'-Converse entailment : if |=b $\rightarrow$ a then a |< b

It is easy to see that A1' implies A1 (by taking a=b=g) and A2' implies A2 (by taking a=b). And the converse is true if the abductive reasoning follows the axioms A1-A5. This two properties have interesting interpretations, which can be also falsified (and in this case the whole set of axioms also falls, by meta-applying A1' to the present theory of abduction !) :

- the first property, as usual, explains how to falsify a hypothesis : if one can abduce an hypothesis from a fact, and if a second fact which is a consequence of both the first fact and the hypothesis is false, then one can no longer abduce the hypothesis.

- the second one expresses an usual property of abduction : if a hypothesis implies some consequence, one can abduce this hypothesis when observing this consequence ; but the converse is not generally true : if one can abduce an hypothesis from a fact, it does not follow that one can deduce the fact from the hypothesis.

1.2. From beliefs to knowledge and the abductive reasoning of an expert

In the previous part, abductive reasoning is based on beliefs, and there are preferences over beliefs. Here, the previous approach is strongly restricted by adding a supplementary axiom for abduction. As a result, there is a unique class of preferences, which is the whole set of worlds : in other words, the previous set K becomes the whole set W. This new situation corresponds to the case in which belief becomes knowledge ; and, as a consequence, abduction is nothing else than the relation reciprocal to deduction.

The new axiom means the transitivity of abduction : if it is possible to abduce a first hypothesis from a fact and if it is possible to abduce a second hypothesis from the first, then it is possible to abduce the second hypothesis from the fact.

A6-Transitivity : if a |< b and b |< g then a |< g

The following representation theorem generalizes a little the one of Flach [Flach, 96], because his axioms are weakened. As already announced, the set K is trivialized as the whole set W, and thus min(B) is nothing else than B.

Theorem 2 (representation theorem) : the abductive reasoning « |< » follows the axioms A1-A6 if and only a |< b iff $B \subseteq A$, i.e. iff b |- a.

Let us remark that, by adding this new axiom, only one abduction relation remains which is exactly reciprocal to the deduction relation. But, this presupposes that beliefs are true, i.e. are knowledge.

Let us come back to the physician's abductive reasoning. Let I be the set of the observable signs and $S=\{s_i\}$ for $i \in$ the set of the elementary propositions corresponding to the presence/absence of the possible signs of possible diseases. Let J the set of possible diseases and $H=\{h_j\}$ for $j \in J$ the set of the elementary propositions corresponding to the presence/absence of the possible diseases (no matter whether the diseases are pure or combined : if it is a combined disease, it will be considered as a different disease). Let W_S the set of possible worlds of signs, W_H the set of possible worlds of diseases and $W= W_S \times W_H$ the set of possible worlds, constructed from the elementary propositions : an elementary world is an interpretation from W into {true, false}. Then the causal relation from hidden diseases towards observable signs can be expressed as logical rules $h_j \rightarrow b_j$ where « b_j » is a proposition in W_S which means that if it is the case that the disease is « j » then necessarily the observable signs satisfy b_j or , equivalently, that the possible world is in B_j.

Now, let us suppose that the physician's abductive reasoning follows A1-A6, i.e. her beliefs are true. Then, following the second representation theorem, when she has the information that the world $w_S \in A$, she can abduce the diseases $J_A = \{j \in J : B_j \subseteq A\}$. More precisely, let $\{A_t\}$ be the information process through time « t » of the physician during an interview with a patient : because information is generally increasing, this sequence of sets $\{A_t\}$ is decreasing. At each stage of this information process, we can define the hypotheses still possible : $J_t = \{j \in J : B_j \subseteq A_t\}$. Because the information is increasing, the sequence A_t is decreasing and the sequence J_t is also decreasing : thus the sequence converges towards a set of possible diagnoses (with zero, one, two or more elements). If finally it is true that all the diseases are known and that there are enough signs (using complementary analyses if necessary) to discriminate each disease from the others ($B_J \cap B_j' = \emptyset$), then there is an information sequence which converges to the right diagnosis for this perfectly competent physician.

Let us next suppose that the physician's abductive reasoning follows only A1-A5. By the first representation theorem, her belief are structured into successive spheres of beliefs with less and less competence from the inner sphere K_0 towards spheres K_1, K_2,..., K_n.. Thus, for this particular physician, her set of diseases J is separated into the sets $J_n = \{j \in J : K_0 * B_j \subseteq K_n\}$ and $J= \cup_n J_n$. Then, with information A_t at time « t » of the appointment, her set of diagnoses $J = \{j \in J : K_0 * B_j \subseteq A_t\}$ has to be separated into the sets of strategies $J_{tn} = \{j \in J_n : K_0 * B_j \subseteq A_t\}$. If J_n becomes empty or the possibility of serious diseases remains true outside J_n then she will probably send the patient to another competent physician.

1.3. Abduction and revision of beliefs

One of the most important general arguments in favor of axioms A1-A5 comes from the literature on revision of beliefs. As stated by the representation theorem, beliefs are then structured into ordered spheres. But the question is then : how it is ontogenetically possible that beliefs be structured in such a way ? A global answer to this question comes from revision belief theory revision [Alchourron & al., 85 ; Gärdenfors, 88; Katsuno & al., 91].

The main question of belief revision theory is how initial belief, represented by a set of worlds K, is modified by the arrival of a message A and becomes the revised belief K*A. The $\in$ method consists in making explicit the axioms for the revision operator K* (which are also falsifiable) and to discuss the deep consequences of the axioms through a representation theorem :

(Axioms for belief revision)

B1-Consistency : if K $\neq$ Ø and A $\neq$ Ø then K*A $\neq$ Ø

B2-Success : K*A $\subseteq$ A

B3-Conservation : if K $\subseteq$ A then K*A=K

B4-Sub-expansion : (K*A) $\cap$ B $\subseteq$ K*(A $\cap$ B)

B5-Super-expansion : if (K*A)  B $\neq$ Ø then K*(A $\cap$ B) $\subseteq$ (K*A) $\cap$ B

Theorem 3 (Representation theorem for belief revision) :

The revision operator « K* » follows the axioms B1-B5 if and only if there is a total preorder relation « $\leq_K$ » on the worlds set W with minimal worlds K such that K*A= min(A) where min(A)={w $\in$ A :w' $\leq$ w for all w' $\in$ A}.

The behavior of the revision operator is easy to understand geometrically. If A and K are compatible, the final belief becomes simply K*A=K $\cap$ A. But the most interesting question is how to restore the consistency (axiom B1) of the beliefs when A and K are incompatible: the solution consists in taking the worlds of A closest to K. Furthermore, the initial system of spheres related to « $\leq_K$ » becomes a new system of spheres related to « $\leq_{K*A}$» with K*A at the center and with peripheral spheres at the intersection of A with the rest of initial spheres.

Now it is clear how abductive reasoning and belief revision are linked together. Abductive reasoning, like deductive reasoning, happens in a static context of beliefs. Revision of beliefs appears to represent the dynamics of beliefs. But both operators, -abduction and revision -, take place in a set of worlds structured by the same total preorder relation.

There is no change in the belief base during the appointment between a physician and a patient. The change happens only when the diagnosis of a specific disease is confirmed by specific experiments to directly characterize the disease. And what the physician has to change is the set B related to this specific disease « j » : more precisely she has to change her hypothesis about this set. It is still an abduction, at a higher level.

Even if it is possible to treat this new kind of abduction in a pure set-theoretic framework, the choice is to shift towards the connectionist paradigm, which has this question at the very core of its framework.

2. Abduction and connectionism

The main point of this part is how an hypothesis about the pattern of a difficult observable concept in the space of easy observable signs can be constructed from positive and negative samples of this concept (for example, a disease). If the observable signs are also propositions, - we will suppose it in the following for the sake of simplicity -, this pattern is a subset of the hypercube $\{0,1\}_n$ defined by the observable signs in number « n ». Also, for the sake of simplicity, we will suppose that the concept is deterministic, i.e. that Nature answers always in the same way at each point of the hypercube.

2.1 What is an hypothesis about the pattern?

The pattern is a subset of the hypercube. Because we have supposed that the concept is a Boolean function, it is well known that the pattern can be represented as a logical formula in a normal form. But this kind of representation is difficult to transform during learning. Here the space of possible worlds is the huge space of all the subsets of the hypercube. This is the main reason for selecting connectionist architectures in order to represent the pattern. The most important *a priori* constraint is that the representation must be generic, i.e. that any pattern can be represented by the architecture.

The most famous architectures are structures based on perceptrons. A perceptron is an hyperplan that separates the hypercube into two subspaces. It is not frequent that only one perceptron is sufficient to separate exactly examples and counterexamples of the same concept. More complicated architectures are generically necessary like the multilayer perceptrons [McClelland & al., 86] or perceptron membranes [Deffuant & al., 96]. The former is well known ; it can represent any pattern, even if there is only one hidden layer, when the number of hidden units is sufficient ; a perceptron membrane is a union of convexes, each one defined by a set of perceptrons : if the set of perceptrons is $\{Q(X, A_i)\}_{i \in I}$ where $Q(x, A_i)$ is linear in X and A_i then the convex is simply the set $C=\{X \in R_n : \wedge_{i \in I} Q(X, A_i) \geq 0\}$; it can be proved that any pattern can be represented as an union of such convexes.

Another representation supposes that the units in the network can perform not only weighted sum of the outputs of the connected units ($\sum$ links) but also weighted sum of different products of these outputs ($\sum \Pi$ links). In this case, the network is any polynom Q(X, A). This approach has been developed more recently under the name « support vector » [Cortes & Vapnik, 95]. Let us remark, in this case, that this network can be thought also as a perceptron because Q is linear in A, the coefficients of the polynom: but this perceptron does not operate directly in the space of X but in the huge larger space where each dimension is a particular product of the components of X. If the number of terms of the polynom is sufficient, all supports can be represented.

The three above kinds of architectures are all generic and able to represent any support. Other architectures have been proposed but it is not useful to describe them for the present purpose. From the general point of view on abduction developed here, it is sufficient to understand (i) that a connectionist network is a non-parametric model Q(X,A) (where A is the coefficient vector associated to a particular architecture of this network) and (ii) that new units can be added to the network (i.e. the non-parametric model with coefficient vector A can be embedded into a larger space of non-parametric models with coefficient vector A') .

Now it is quite clear what is an hypothesis about the support. There are two parts in such an hypothesis : the first part is an hypothesis about the structure of the network ; and the second part is an hypothesis about the

value of the coefficients. The space of hypotheses is a tree (S, G) of spaces : each space $s \in S$ is isomorph to an R_m if there is « m » coefficients to define the network structure in this space and is embedded in the directly accessible spaces G_s in this tree.

2.2. Admissible hypotheses of support

The non-parametric model Q(X,A) in space $s \in S$ gives true or false answers for X according to whether $Q_s(X,A)$ is positive or negative. Thus, a basic remark is relevant for all the following : by a kind of duality, each sample z={X, y} with $y \in$ {true, false} as answered by Nature introduces in the current space of hypotheses a constraint :

H_{sz} ={A $\in R_m$: $Q_s(X,A) \geq 0$ if y=true and $Q_s(X,A)<0$ if y=false}

In the case of a perceptron or of support vector, $Q_s(X,A)$ is linear in A : this constraint has the geometrical form of an hyperplan that separates the whole space in two subspaces with only one permitted. The same geometrical intuition is true for the general case, but the separation is more complicated than an hyperplan.

Given an available set Z of samples, the set of admissible hypotheses is easy to define, at least in principle : $H_{sZ} = \cap_{z \in Z} H_{sz}$. When the number of samples increases, each set of admissible hypotheses H_{sZ} decreases for all $s \in S$ and can become empty. This qualitative analysis can be interpreted in two ways, in the belief revision context and in the abduction context and fit very well in both frameworks.

The first way, related to belief revision, is convenient for understanding learning process, as a diachronic process : at a given stage of learning with the set of samples Z, the network belongs to $s \in S$; if a new sample z arrives and is compatible with H_{sz} then the network remains in s ; if it is not the case, it is necessary that the network leaves the space s towards one of its directly accessible spaces G_s in S. This dynamics is conform, at least qualitatively, with the revision of belief : H_{sZ} plays the role of an initial belief, H_{sz} the role of a message ; if $H_{sZ} \cup_z = H_{sZ} \cap H_{sz} \neq \emptyset$, then the initial belief and the message are compatible and $H_{sZ'}$ with $Z'=Z \cup z$ is the revised belief; if it is not the case, the first extension of spheres of beliefs concerns the directly accessible spaces G_s in S ; if the consistency cannot be restored in the directly accessible spaces, the second extension of spheres has to move to the next accessible spaces. The fact that the extension stops at the first sphere which restores the consistency can be interpreted as a token of the Occam razor procedure. Let us notice that an extension can lead to more than one directly accessible space : in this case, an irreversible bifurcation necessarily occurs, which influences all future learning. Thus each learning trajectory of a network can move in a lot of different spaces, depending on the bifurcations and on the order in the presentation of the samples.

The second way, related to abductive reasoning, is more convenient to understand in a synchronic way all the possible states of a population of networks, when they have encountered the same set of samples, possibly in different orders. Indeed, given a set of samples Z (given in whichever order), abductive reasoning eliminates all the spaces $s \in S$ with $H_{sZ} = \emptyset$ and keeps only the minimal elements of the tree of spaces with $H_{sZ} \neq \emptyset$, i.e. also the Occam razor :

(i) the admissible spaces are S(Z)=min(S'$_Z$) where S'$_Z$ ={ $s \in S$ with $H_{sZ} \neq \emptyset$}

(ii) the admissible hypothesis are $\sum \sum (Z) = \sum_{s \in S(Z)} H_{sZ}$

This preference for minimal, parsimonious hypotheses is exactly what is specified by the representation theorem. Furthermore, because no minimal admissible state is discarded, all the particular learning networks is in one of these states $\sum (Z)$.

2.3. Convergence of belief to knowledge

Because the concept we have considered is deterministic, and because learning restores always consistency, the convergence of the network towards the concept is warranted when the set of samples is increasing until it covers all the sample space. That means that the final belief is semantically the same in all networks, even if the different final networks are not in the same spaces or are in the same space but not with the same coefficient vector. The situation here is analog to what happens with logical formulae representing a concept : many formulae can represent semantically the same concept.

3. Abduction and constructivism

As stated in the introduction, the point of view of constructivism leads to a change of the criterium of success : what is asked to Nature is not whether a proposition is true or false but whether it leads to life or death. In other words, the criterium is viability rather than validity. In both cases, anticipations are required ; there is in fact no opposition between the two attitudes : if valid anticipations can be done, the probability to remain in the domain of viability increases. Nevertheless, the point of view of constructivism brings the evolutionist criterium of viability to the fore.

The main difficulty with the criterium of viability is to be a criterium with a long term horizon. It is very hard to anticipate what an immediate action will change for the long term viability. For a young individual living inside a society of individuals of the same type, one excellent strategy is to imitate the behavior of the older individuals : her probability to live older will increase (at least in stable environments). More generally, an excellent strategy for a novice wanting to acquire a know-how consists in imitating the know-how of experts. The only presupposition is that she can recognize who is expert.

This explains why the model of abduction presented below concerns essentially *mimetism*. Other models of how an expert develops further her know-how on the basis of her experience is left outside of the scope of this paper, even if it is important. As one will see, the model is quite similar to the preceding one. But the meaning and what is to be learned is different . There are two differences : the first one is introduction of categorization ; the second consists in constructing a utility function to measure success, whether it means viability or improved know-how.

3.1. How does know-how emerge by mimetism ?

In order to make more explicit how know-how emerges by mimetism, let us first consider a novice chess player, before trying to generalize to more complex situations. She knows the rules of the game, i.e. the graph (X,G) where X is the space of the possible situations and G(X) is the directly accessible situations from the current situation X. She can observe games of a chess master. What she has to construct is an accurate evaluation of whether a situation is better than another for reaching success. Indeed, if she has such an evaluation, choosing the next move becomes strongly abductive : take the move in G(X) which has the best evaluation.

At first, such an evaluation is a very difficult anticipation, because the consequences of a move take place only in the long term, at the end of the game. The key question is to learn such an evaluation by observing a chess master (or many). There is essentially one principle that is useful in the observation of the strategies of a master, *the humanity principle* : it is an interpretative principle which attributes rationality to others' behavior. This principle gives a huge quantity of information, because each move of the master has to be interpreted as

leading to a better situation than all other accessible situations. What this principle allows the novice to learn is a model of the master's preferences. If the master's preferences follow the classical axioms of the rational agent in microeconomy theory, then her preferences can be represented by a utility function Q(X).

A new difficulty now arises because such a function can be very complicated : it is well known that the dependence of Q from x takes into account very subtle combinations of pieces, i.e. patterns of pieces. But there are one piece patterns, two pieces patterns,… By adding new refinements of patterns, i.e. by augmenting the number of interacting pieces or secondary features on the chess map, we obtain a hierarchy (S,G) of « description spaces » where each new specialized pattern adds a new dimension. When a space « s » is chosen in S, the situation is represented by a vector Y of present/absent patterns which depend functionally on the situation X : in other words $Y=Y_s(X)$. The framework is exactly the same as for support vector in the part 2 : as a corollary, what is to be searched is a function $Q_s(Y,A)$ that depends linearly on coefficient vector A.

Before continuing the modeling, let us consider the preceding argument in order to generalize it for any relation between novices and experts. In fact there is nothing specific in the argument based on the principle of humanity and on the axioms about preferences. The main counterargument concerns the kind of situations that occur in chess games : a chess map is a microworld ; thus the novice can easily refine pattems by herself. It is not the case for the real world : here it is very difficult for a novice to extract patterns relevant in a given situation. Her task is one of categorizing [Rosch, 78] and of refining more and more her basic categories. One of the main help that an expert can bring to a novice is to make explicit the patterns/prototypes/categories which she uses to characterize a situation in a relevant way. It is enough to think just a little while of the complexity of the world and of human cognition to understand that categorization is probably the main part of human cognition. But, nevertheless, at least in principle, the process of categorization can begin with the expert and be pursued by the novice herself. The argument resists in the general case.

There is another more radical counterargument : it concerns a class of situations where the novice can't describe the situation that would occur if another action was performed by the expert ; in other words, the novice ignores the above counterfactual relation G(X) on the situations. In such class of situations, the novice can only model the preference of the expert for some action in a situation versus the other actions. What the novice tries to imitate is no more the utility the expert attribute to situations but the utility she attributes to actions conditionally to a situation. This restrictive case is not developed here but it is based on the same principles (humanity principle and axioms of rational behavior) and leads to a similar treatment in what follows.

3.2.Admissible hypotheses of a utility function

Let us summarize the previous discussion : the expert is choosing a path Z in a graph (X,G), where X is the space of situations and G is the relation of accessibility of new situations from the present situation. The situation can be described by choosing a space of descriptions « s » belonging to a set S of such spaces : in this particular space, the description of a situation depends functionally on this situation : $Y= Y_s(X)$; this dependency can be arbitrarily complicated without any difficulty for what follows. One should now define what is an admissible non-parametric model $U_s(Y_s(X),A)$ for the utility function of the expert in this space of descriptions.

What is implicitly revealed by each choice is very simple. Let z=(X,X') be one element of the path. It is sufficient to write that x' is preferred to all other accessible situations belonging to $X'' \in G(X)$:

$H_{sz} = \{A \in R_m : U_s(Y_s(X'),A) \geq U_s(Y_s(X''),A)$ for all $X'' \in G(X)\}$

Given a path Z of samples, the set of admissible hypotheses on utility function is easy to define: $H_{sZ} = \cap_{z \in Z} H_{sz}$. When the length of the path increases, each set of admissible hypotheses H_{sZ} decreases for all $s \in S$ and can vanish. And the novice has to shift towards a more sophisticated space of descriptions in order to restore the consistency about what she models as the utility function of the expert.

All the remaining line of reasoning is exactly the same as in the connectionist part. Each novice can find a particular way to refine her models of an expert through her progressive study of this path. All novices at the same given stage Z of study have one of the admissible models $\sum$ (Z) defined by the same relations (i) and (ii) of part 2.2, as a result of abductive reasoning. When the path tends to an infinite length, all models converge semantically (but not syntactically).

If there are many experts, they generally have different preferences and this is essential in an evolutionary point of view. And novices have preferences (or simply opportunities) for imitating such or such an expert. As mediated by the imitation of utility functions of the experts, expert preferences diffuse more or less asymmetrically through the population of novices, which can then explore original paths. Such explorations introduce variations in the preferences and lead some novices to become experts and to play the converse role of imitatees.

4.Conclusion

As stated in the introduction, the purpose was to give to the concept of abduction a status of the same importance as to deduction. The representation theorem is the main proposition providing a unification of the meanings of abduction. This theorem has two equivalent formulations. The first delivers a set of axioms ; it is convenient for an expert or a scientist. The second offers a geometrical understanding of how abduction and deduction are linked through a system of spheres defined by a total preorder on the possible worlds, which implies a preference relation on the set of beliefs : any hypothesis contains a preferred part, its corresponding parsimonious hypothesis ; all happens as if the hypothesis interacted with the system of beliefs only through its parsimonious hypothesis during abductive inferences ; and for parsimonious hypotheses, abduction is exactly the relation reciprocal to deduction. In other words, abductive inferences can be performed only with parsimonious hypotheses as reciprocal to deductive inferences.

Like deduction, abduction is a synchronic relation ; there is no change in the system of beliefs through deduction or abduction. Change occurs only when a new message is issued by Nature : this moment of induction is followed by a revision of beliefs which consists in selecting the preferred part of the message. Thus, if beliefs are structured by a relation of preference, abduction, deduction and induction have together a simple and natural meaning.

The axioms can be seen as normative or as descriptive. They can have a normative sense for some human activity like scientific research and other ones (like in the research of a culprit, when they become more or less explicit conventions on what should be acceptable abductive reasoning. They can be descriptive and then can be falsified by psychological experiments : if it was the case, it would be necessary to weaken some axioms, i.e. to enlarge the theory of abduction.

If the axiomatic approach of abduction is the right approach within the cognitivist paradigm, the geometric approach is the right approach of abduction for the connectionist and constructivist paradigms of cognition. Indeed the geometric approach allows to understand how hypotheses of higher levels –prototheories- can

emerge from sensory experience. This geometric approach has been discussed for non-parametric models associated to a hierarchy of spaces of larger and larger networks. A natural hierarchy of hypotheses appears with the same structure as a system of spheres. Parsimonious and consistent hypotheses are exactly those obtained by selecting in the hierarchy of spaces the minimal spaces having at least one model consistent with the observed facts. The geometrical approach seems to be truly promising for the connectionist and the constructivist paradigm, even if a few modes of learning have been discussed. Enlarging the discussion to any kind of learning is a very important challenge for a better understanding of abduction, as the « mode of reasoning of the living », as coined by Peirce.

Notes :

1. See the forthcoming CREA-report of Bourgine & Walliser to appear.

References :

Alchourron C.E., P. Gärdenfors & D. Makinson, 1985, On the logic of theory of change : partial meet contraction and revision functions, Journal of Symbolic Logic, 50, 510-530.

Anderson, Douglas R, 1986, The Evolution of Peirce ‘s Concept of Abduction, Transactions of the Charles S. Peirce Society 22, no 2, 145-164.

Aubin J.-P. , 1991 : Viability Theory, Birkhäuser.

Bochereau L., Bourgine P., 1990, Extraction of semantic features and logical rules from a multilayer neural network, Proceedings IJCNN 90, Washington,DC.

Bochereau L., Bourgine P. and Deffuant G., 1990, "Equivalence between Connectionist Classifiers and Logical Classifiers", in Lecture Notes in Physics 368, Springer Verlag, 1990, pp. 351-363.

Bourgine P., 1991, Heuristique et Abduction, PhD, Cahiers du LIUC n° 95, Université de Caen.

Bourgine P., F. Varela, 1992, Towards a practice of autonomous system, in Towards a practice of autonomous system, F.Varela & P.Bourgine (ed). MIT Press/Bradford Books.pp 3-10.

Bourgine P., Walliser B.,1992, Economics and Cognitive Science, Pergamon Press.

Burks, A , 1946, Peirce Theory of Abduction, Philosophy of Science 13, p.301-306.

Cortes C., V.Vapnik, 1995, Support-Vector Networks, Machine Learning, 20, 273-297.

Fann, K.T., 1970, Peirce’s theory of Abduction, Martinus Nijhoff, the Hague.

Deffuant, G., Fuchs, T., Monneret, E., Bourgine, P., and Varela, F., 1996, Semi-algebraic organisms: the morphodynamical network perspective in Artificial Life 2:157-179.

Flach P.A.,1996, Abduction and Induction : Syllogistic and Inferential Perspectives, Workshop on abduction and induction, Budapest.

Funahashi K.I. 1989. On the approximate realization of continuous mappings by neural networks. Neural Networks, Vol.2, pp. 183-192.

Gärdenfors Peter, 1988. Knowledge in flux. M.I.T. / Bradford Books, Cambridge, MA.

Katsuno A.&A. Mendelzon, 1991, Propositional knowledge base revision and minimal change, Artificial Intelligence, 52, 263-294.

Kohonen T., 1984. Self-Organization and Associative Memory. Springer Verlag.

Kraus S., D. Lehmann & M. Magidor, 1990, Nonmonotonic reasoning, preferential models and cumulative logics, *Art. Intel.* **44** : 167-207.

McClelland James L., David E. Rumelhart, 1986, Parallel Distributed Processing : Exploration in the Microstructure of Cognition. MIT/Bradford Book, Cambridge, MA.

Peirce, C.S. The collected Papers., Hartshorne, Weiss & Burks (eds). 8 Vls. Harward University Press, Cambridge, MA .

Rosh,E., 1978, Principles of Categorization, in Cognition and Categorization, ed. E.Rosh and B.B.Lloyd, Lawrence Erlbaum, Hillsdalle, N.J., 27-48.

Santaella, L., 1997, The development of Peirce's three types of reasoning : Abduction, Deduction, and Induction, International Semiotic Congress, Mexico, June 97.

Thagard, Paul R., 1977, The Unity of Peirce's Theory of Hypothesis , Transactions of the Charles S.Peirce Society 13, no2, 112-123.

Thagard, Paul R.,1981, Peirce on Hypothesis and Abduction, Proceedings of the C.S. Peirce Bicentennial International Congress, K.L. Ketner & al. (eds.), 271-274.

Walliser B., D. Zwirn, 1998, Probabilistic Belief Revision Principles, ENPC Technical Report, Paris.