Abduction, Reason, and Science

Processes of Discovery and Explanation

Lorenzo Magnani

University of Pavia
Pavia, Italy, and
Georgia Institute of Technology
Atlanta, Georgia

Kluwer Academic / Plenum Publishers
New York, Boston, Dordrecht, London, Moscow

Science does not rest upon solid bedrock. The bold structure of its theories rises, as it were, above a swamp. It is like a building erected on piles. The piles are driven down from above into the swamp, but not down to any natural or "given" base; and if we stop driving the piles deeper, it is not because we have reached firm ground. We simply stop when we are satisfied that the piles are firm enough to carry the structure, at least for the time being.

Karl R. Popper, *The Logic of Scientific Discovery*

Preface

This volume explores abduction (inference to explanatory hypotheses), an important but neglected topic in scientific reasoning. My aim is to integrate philosophical, cognitive, and computational issues, while also discussing some cases of reasoning in science and medicine. The main thesis is that abduction is a significant kind of scientific reasoning, helpful in delineating the first principles of a new theory of science.

The status of abduction is very controversial. When dealing with abductive reasoning misinterpretations and equivocations are common. What are the differences between abduction and induction? What are the differences between abduction and the well-known hypothetico-deductive method? What did Peirce mean when he considered abduction a kind of inference? Does abduction involve only the generation of hypotheses or their evaluation too? Are the criteria for the best explanation in abductive reasoning epistemic, or pragmatic, or both? How many kinds of abduction are there?

The book aims to increase knowledge about creative and expert inferences. The study of these high-level methods of abductive reasoning is situated at the crossroads of philosophy, epistemology, artificial intelligence, cognitive psychology, and logic; that is, at the heart of cognitive science. Philosophers of science in the twentieth century have traditionally distinguished between the inferential processes active in the logic of discovery and the ones active in logic of justification. Most have concluded that no logic of creative processes exists and, moreover, that a rational model of discovery is impossible. In short, scientific creative inferences are irrational and there is no "reasoning" to hypotheses (chapter 1). On the other hand, some research in the area of artificial intelligence has shown that methods for discovery could be found that are computationally adequate for rediscovering - or discovering for the first time - empirical or theoretical laws and theorems (chapter 2). Moreover, the study of diagnostic (chapter 4), visual, spatial, analogical, and temporal reasoning (chapter 5) has demonstrated that there are many ways of performing intelligent and creative reasoning that cannot be described with only the help of classical logic. However, non-standard

logic has shown how we can provide rigorous formal models of many kinds of abductive reasoning such as the ones involved in defeasible and uncertain inferences (chapter 2).

Contradictions and inconsistencies are fundamental in abductive reasoning, and abductive reasoning is appropriate for "governing" inconsistencies. Many ways of governing inconsistencies will be considered (chapter 6), from the methods activated in diagnostic settings and consistency-based models to the typical ones embedded in some forms of creative reasoning, from the interpretations in terms of conflicts and competitions to the actions performed on empirical and conceptual anomalies, from the question of generating inconsistencies by radical innovation to the connectionist treatment of coherence.

The interdisciplinary character of abduction is central and its fertility in various areas of research evident. The book also addresses the central epistemological question of hypothesis withdrawal in science by discussing historical cases (chapter 7), where abductive inferences exhibit their most appealing cognitive virtues. Abduction is also useful in describing the different roles played by the various kinds of medical reasoning, from the point of view both of human agents and of computational programs that perform medical tasks such as diagnosis (chapter 4).

Finally, an interesting and neglected point of contention about human reasoning is whether or not concrete manipulations of external objects influence the generation of hypotheses, for example in science. I will delineate the first features of what I call manipulative abduction, showing how we can find methods of constructivity in scientific and everyday reasoning based on external models and "epistemic mediators" (chapter 3).

During the period in which this book was written, I was Visiting Professor of Philosophy of Science at Georgia Institute of Technology, Atlanta, which provided an excellent working environment. I am grateful to my colleagues there at the College of Computing and School of Public Policy for their helpful suggestions and much more. For valuable comments on a previous draft I am particularly grateful to Ronald Giere, David Gooding, Kenneth Knoespel, Nancy Nersessian, Paul Thagard, and two anonymous referees. Special thanks to Nancy Nersessian and Paul Thagard, who in the last ten years played a significant role in shaping my ideas and in helping me to focus and articulate my views.

The research related to this volume was supported by grants from the Italian Ministry of University, University of Pavia, CNR (Centro Nazionale delle Ricerche), CARIPLO (Cassa di Risparmio delle Provincie Lombarde), and Ivan Allen College (Georgia Institute of Technology). The preparation of the volume would not have been possible without the contribution of resources and facilities of the Computational Philosophy Laboratory (Depart-

ment of Philosophy, University of Pavia), and of Georgia Institute of Technology.

This project was conceived as a whole, but as it developed various parts have become articles, which have now been excerpted, revised, and integrated into the current text. I am grateful to the respective publishers for permission to include portions of previously published articles.

Magnani, L., forthcoming, Creative abduction and hypothesis withdrawal in science, in: *Methodological Aspects of Discovery and Creativity*, J. Meheus and T. Nickles, eds., Kluwer, Dordrecht.

Magnani, L., 1999, Creations and discoveries in science: the role of abductive reasoning, in: *Human and Machine Perception II: Emergence, Attention, and Creativity*, V. Cantoni, V. Di Gesù, A. Setti, and A. Tegolo, eds., Kluwer Academic/Plenum Publishers, New York, pp. 137-149.

Magnani, L., 1999, Model-based creative abduction, in: *Model-Based Reasoning in Scientific Discovery*, L. Magnani, N. J. Nersessian, and P. Thagard, eds., Kluwer Academic/Plenum Publishers, New York, pp. 219-238.

Magnani, L., 1999, Withdrawing unfalsifiable hypotheses, *Foundations of Science* 4(2):133-153.

Magnani, L. 1997, Basic science reasoning and clinical reasoning intertwined: epistemological analysis and consequences for medical education, *Advances in Health Sciences Education* 2: 115-130.

Magnani, L., Chella, A., and da Fontoura Costa, L., 1999, Symbolism and connectionism paradigms, in: *Human and Machine Perception II: Emergence, Attention, and Creativity*, V. Cantoni, V. Di Gesù, A. Setti, and A. Tegolo, eds., Kluwer Academic/Plenum Publishers, New York, pp. 185-194.

Magnani, L. Civita, S., and Previde Massara, G., 1994, Visual cognition and cognitive modeling, in: *Human and Machine Vision: Analogies and Divergences*, V. Cantoni, ed., Plenum Press, New York, pp. 229-243.

Lorenzo Magnani
University of Pavia, Pavia, Italy
Georgia Institute of Technology, Atlanta, GA,

Chapter 1

Hypothesis Generation

1. REMINISCENCE, TACIT KNOWLEDGE, SCHEMATISM

The themes introduced in this chapter illustrate some important aspects of hypothesis generation central to correctly posing the problem of abduction. In the history of philosophy there are at least three important ways for designing the role of hypothesis generation, always considered in the perspective of problem solving performances. All aim at demonstrating that the activity of generating hypotheses is paradoxical, either illusory or obscure, implicit, and not analyzable.

Plato's doctrine of *reminiscence* can be looked at from the point of view of an epistemological argument about the paradoxical concept of "problem-solving": in order to solve a problem one must in some sense already know the answer, there is no real generation of hypotheses, only recollection of them. The activity of Kantian *schematism* is implicit too, resulting from imagination and completely unknowable as regards its ways of working, empty, and devoid of any possibility of being rationally analyzed. It is an activity of tacit knowledge, "an art concealed in the depths of the human soul, whose real modes of activity nature is hardly likely ever to allow us to discover, and to have open to our gaze". In his turn Polanyi thinks that if all knowledge is explicit and capable of being clearly stated, then we cannot know a problem or look for its solution; if problems nevertheless exist, and discoveries can be made by solving them, we can know things that we cannot express: consequently, the role of so-called *tacit knowledge* "the intimation of something hidden, which we may yet discover" is central.

It is very useful to focus our attention on an ancient philosophical story, which referring to the Platonic doctrine of recollection in the famous *Meno* dialogue; the story facilitates various aims: 1. to illustrate the relevance of the activity of guessing hypotheses, dominant in *abductive reasoning*, as we will see in the following chapter; 2. to explain the so-called "generate and test" model (cf. the following section), proposed by Herbert Simon in the Sixties, that leads to the intellectual atmosphere of problem-solving (Simon, 1965, Newell and Simon 1963, 1977; Newell, 1982, Newell, 1990), and 3. to initiate the reader into the multiple aspects of the concept of abduction: since Peirce's landmark definition (cf. chapter 2, section 1), abduction has been convincingly modeled as a process of generating and testing[1].

In ordinary geometrical proofs auxiliary *constructions* are present in terms of "conveniently chosen" figures and diagrams where strategic moves are intertwined with deduction (Hintikka and Remes, 1974; Hintikka, 1998). The system of reasoning exhibits a dual character: "hypothetical" and deductive. This dual character is also illustrated by the method of analysis and synthesis in Greek geometry, one of the most important ideas in the history of heuristic reasoning. In deduction, that is *synthesis*, reasoning proceeds from causes to their effects. In theoretical analysis, reasoning goes backward from theorems to axioms - from effects to causes - from which they deductively follow. In the so-called problematical *analysis*, which attempts to solve geometrical problems, the desired target (that defines the so-called "model-figure" as a "construction") is assumed to be given, and the reasoning again goes backward looking for possible constructions from which the sought target results. The story of *Meno* dialogue will illustrate the role of these strategical "analytical" moves and their importance in hypothesis generation[2].

Commenting some modern reinterpretations of the concept of analysis Niiniluoto says:

> Hintikka and Remes (1974) make important objections to the propositional interpretations of analysis. One of their observations is that theorems in geometry are typically general statements (e.g., "for all triangles, the sum of their angles equal 180 degrees") or universal-existential statements (e.g., "for all geometrical figures x, if x is a square, then x will have two diagonals and these diagonals bisect each other"). Proof of such general implications proceeds through their instantiations and by at-

[1] The *Meno* story is also connected to the centrality of the concept of problem-solving in teaching and learning, illustrated in chapter 4, where I will deal with the problem of medical education and knowledge-based systems (KBSs) in medicine.

[2] In chapter 2, section 4, these moves of problematic analysis will be characterized as abductive. See also in chapter 2, footnote 25, the Newtonian use of analysis and synthesis to characterize the method of the new physics.

tempting to derive the consequent from the antecedent by suitable axioms or rules of inference (Niiniluoto, 1999, p. 244).

Plato's *Meno* is a fascinating dialogue about whether virtue can be taught (Turner, 1989). At the end of the dialogue Socrates states that if virtue is teachable then it could be taught either by the Sophists or by virtuous men. Socrates however illustrates that many virtuous men had taught virtue to their sons but had failed to make them "virtuous". Nor should we expect the Sophists to be able to teach virtue; they only make men clever orators. Socrates concludes that virtue is not teachable: it is divinely bestowed. The slave boy in the dialogue is involved in a "proof" that serves Socrates to demonstrate that "we know more than we can tell", which is the subproblem of the dialogue itself which is about finding something about which we know nothing at all. On this basis, Socrates concludes that "research and learning are wholly recollection" (Plato, 1977, 81 d, p. 303). The slave boy will be able to "recollect" a conclusion equivalent to the Pythagorean theorem from examples, in terms of constructions ("model-figures"), and appropriate questions.

I will closely follow the Platonic text to focus the attention of the reader on the particular philosophical use of expressions like "to know", "to suppose", "contradiction", "doubt", "to learn", "to teach", but also to re-echo the methodological atmosphere given by Plato in the dialogue.

Socrates draws in the sand a square divided into four equal squares (Figure 1) and establishes 1) that the boy cannot correctly answer the question, of how much larger the sides of a square with double the area of another square will be, and - it is postulated that the boy does not know anything about geometry - 2) that the boy thinks he knows that if a square has twice the area the sides will also be double. In response to Socrates' hypothesis that each side of the square ABCD is two feet long, the boy correctly answers that the whole square has a "space" of four. Then Socrates:

> SOC. And might there not be another figure twice the size of this, but of the same sort, with all its sides equal like this one? BOY. Yes. SOC. Then, how many feet will it be? BOY. Eight. SOC. Come now, try and tell me how long will each side of that figure be. This one is two feet long: what will be the size of the other which is double in size? BOY. Clearly, Socrates, double. SOC. Do you observe, Meno, that I am not teaching anything, but merely asking him each time? And now he supposes that he knows about the line required to make a figure of eight square feet; or do you not think he does? MEN. I do. SOC. Well, Does he know? MEN. Certainly not. SOC. He just supposes it, from the double size required? MEN. Yes. (Plato, 1977, 82 d-e, p. 307).

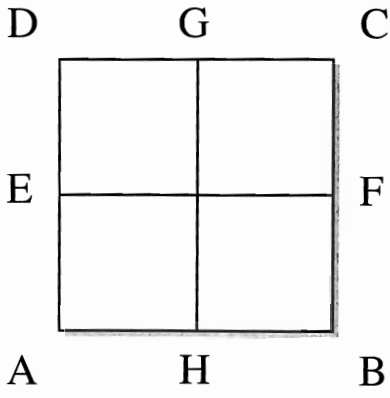

Figure 1. The given square.

At this point Socrates leads the boy through a series of inferences, each of which the boy could either "tell" or simply assent to in response to Socrates' "questions". Socrates shows that by prolonging the sides of the given square with lines of equal length we obtain a square AILM (Figure 2) which is simply the one the boy thinks is an eight-foot figure.

> SOC. And here, contained in it, have we not four squares, each of which is equal this space of four feet? BOY. Yes. SOC. Then, how large is the whole? Four times that space, is it not? BOY. It must be. SOC. And is four times equal to double? BOY. No, to be sure. SOC. But how much is it? BOY. Fourfold. SOC. Thus, from the double-sized line, boy, we get a space not of double, but of fourfold size. BOY. That is true. SOC. And if it is four times four, it is sixteen, is it not? BOY. Yes. (Plato, 1977, 83 b-c, p. 309).

Continuing with his dialogic method, Socrates the dialectic, scrupulous and pitiless, always engaged in delineating definitions, confuting the false ideas and clarifying the confused ones, is leading the boy to find the side of a square which has a eight-feet area, that evidently will be less than four and more than two. In response to the slave's hypothesis that the desired square would have an area of three feet Socrates obviously shows this affirmation to be in contradiction with the evidence that this case gives rise to a nine-foot figure (APRS). Socrates concludes:

> So we fail to get our eight-foot figure from this three-foot line. BOY. Yes, indeed. SOC. But from what line shall we get it? Try and tell us exactly; and if you would rather not reckon it out, just show what line it is. BOY. Well, on my word, I for one do not know (Plato, 1977, 83 e, 84 a, p. 313).

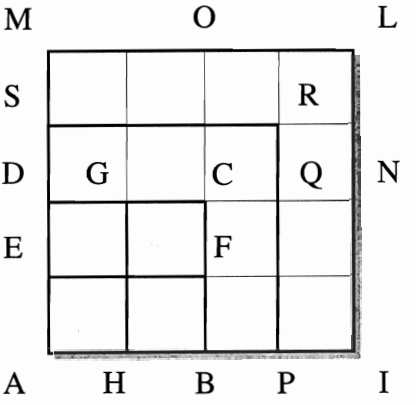

Figure 2. Construction - I.

At this point Socrates turns to Meno and shows him the "progress" the boy "has already made in his recollection", reaching that "doubt" which he considers the primordial philosophical condition. Socrates observes:

> SOC. At first he did not know what is the line that forms the figure of eight feet and he does not know even now: but at any rate he thought he knew then, and confidently answered as though he knew, and was aware of no difficulty; whereas now he feels the difficulty he is in, and besides not knowing does not think he knows. [...] And is he not better off in respect of the matter which he did not know? [...] Now, by causing him to doubt and giving him the torpedo's shock, have we done him any harm? MEN. I think not. SOC. And we have certainly given him some assistance, it would seem, towards finding out the truth of the matter: for now he will push on in the search gladly, as lacking knowledge; [...]. Now do you imagine he would have attempted to inquire or learn what he thought he knew, when he did not know it, until he had been reduced to the perplexity of realizing that he did not know, and had felt a craving to know? [...] Now you should note how, as a result of this perplexity, he will go and discover something by joint inquiry with me, while I merely ask questions and do not teach him; and be on the watch to see if at any point you find me reaching him or expounding to him, instead of questioning him on his opinions (Plato, 1977, 84 a-d, pp. 313-315).

Then, drawing three squares one after the other (Figure 3), Socrates leads the boy to analyze the square (BDON) of side DB (the diagonal) having twice the area of the square given at the beginning of the dialogue, that is of four feet:

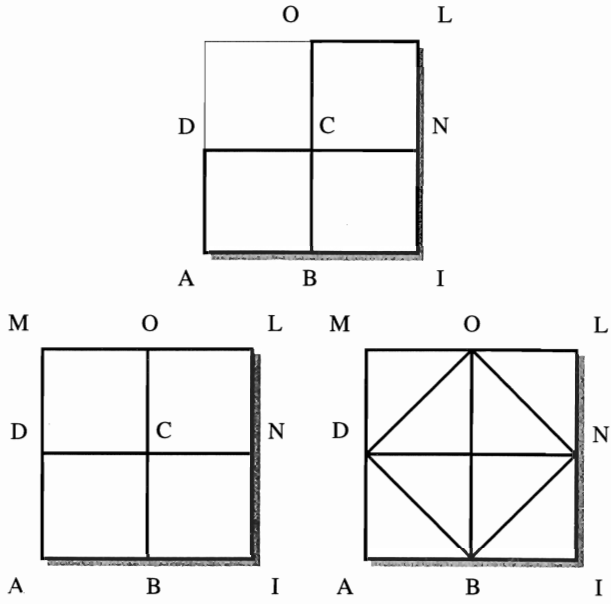

Figure 3. Constructions II-III-IV.

SOC. Tell me boy: here we have a square of four feet (ABCD)? Have we not? You understand? BOY Yes. SOC. And here we have another square (BICN) equal to it? BOY. Yes. SOC. And here a third (CNLO), equal to either of them? BOY. Yes. SOC. Now shall we fill up this vacant space (DCOM) in the corner? BOY. By all means. SOC. So here we must have four equal spaces? Yes. SOC. Well now. How many times larger is this whole space (AILM) than this other? BOY. Four times. SOC. But it was to have been only twice, you remember? BOY. To be sure. SOC. And does this line, drawn from corner to corner, cut in two each of these spaces? BOY. Yes. SOC. And have we here four equal lines containing this space? BOY. We have. SOC. Now consider how large this space is. BOY. I do not understand. SOC. Has not each of the inside lines cut off half of each of these four spaces? BOY. Yes. SOC. And how many spaces of that size are there in this part (BDON)? BOY. Four. SOC. And how many in this (ABCD)? BOY. Two. SOC. And four is how many times two? BOY. Twice. SOC. And how many feet is this space (BDON)? BOY. Eight feet. SOC. From what line do we get this figure? BOY. From this (DB). SOC. From the line drawn corner-wise across the four-foot figure (ABCD)? BOY. Yes. SOC. The professors [*sophists*] call it (DB) the diagonal: so if the diagonal is its name, then according to you, Meno's boy, the double space is the square of the diagonal. BOY. Yes, certainly it is, Socrates (Plato, 1977, 84 d-e, 85 a-b, pp. 317-319).

Socrates remarks that all the opinions expressed by the boy derive from his own thought:

> SOC. So, that he who does not know about any matters, whatever they be, may have true opinions on such matters, about which he knows nothing? [...] And at this moment those opinions have just been stirred up in him, like a dream; [...]. Without anyone having taught him and only through questions put to him, he will understand, recovering the knowledge out of himself? [...] And is not this recovery of knowledge in himself and by himself recollection? [...] Or has someone taught him geometry? You see he can do the same as this with all geometry and every branch of knowledge. Now can anyone have taught him all this? You ought surely to know especially as he was born and bred in your house. MEN. Well, I know that no one has ever taught him. SOC. And has he these opinions or has he not? MEN. He must have them, Socrates, evidently. And if he did not acquire them in this present life, is it not obvious at once that he had them and learnt them during some other time? MEN. Apparently. SOC. And this must have been the time when he was not a human being? MEN. Yes. SOC. So if in both of these periods - when he was and was not a human being - he has had true opinions in him which have only to be awakened by questioning to become knowledge, his soul must have had this cognisance throughout all time? For clearly he has always either been or not been a human being (Plato, 1977, 85 c-e, 86 a, pp. 319-321).

Exploring the secrets of geometry, in the Pythagorean atmosphere of the infinite succession of lives that characterize immortal souls, Plato formulates the ancient theory of true opinions[3] and recollection. The true opinion is given by recollection and science is the system of true opinions when related by the activity of reasoning and thereby made permanent and definitive. "Constructing" the figures, Socrates the dialectic leads the young slave to discover by himself the geometrical truths he already possesses in his spirit. The slave's experience directly assists philosophy and leads us to the classic scenario of the doctrine of reminiscence. The story of Socrates and Meno's slave is the *narrative* that illustrates this famous philosophical theory.

The problem is related to the so-called *Meno* paradox, stated by Plato[4] in the dialogue and discussed by Simon (1976) (see the following section), and

[3] On the difference between knowledge and mere true opinion, and other methodological and epistemological considerations on a modern interpretation of Plato's *Meno*, cf. Glymour (1992).

[4] "SOC. Do you see what a captious argument you are introducing - that, forsooth, a man cannot inquire either about what he knows or about what he does not know? For he cannot inquire about what he knows, because he knows it, and in that case is in no need of in-

to the issue of *tacit knowledge* which was introduced by Polanyi (Polanyi, 1966). Indeed, the story of *Meno*'s slave can be looked at from the point of view of an epistemological argument about the paradoxical concept of "problem-solving" (Bruner, Goodnow, and Austin, 1956; Polya, 1957).

Polanyi thinks the *Meno* story shows that if all knowledge is explicit, i.e., capable of being clearly stated, then we cannot know a problem or look for its solution. It also shows that if problems nevertheless exist, and discoveries can be made by solving them, we can know things that we cannot express: "[...] to search for the solution of a problem is an absurdity; for either you know what you are looking for, and then there is no problem; or you do not know what you are looking for, and then you cannot expect to find anything" (Polanyi, 1966, p. 22). As stated above, Plato's solution of this epistemological impasse is the very classic philosophical scenario of the doctrine of reminiscence: Socrates' teaching is in reality leading the slave to discover the knowledge he already possesses in his spirit.

Following Polanyi's interpretation the geometrical dialogue should not be related to the doctrine of reminiscence: it is the concept of "problem" that characterizes the whole story, where the role of the so-called *tacit dimension* is central. Therefore, the *Meno* story becomes a psychological-epistemological one. Examples of "tacit knowledge" can be found in *Gestalt* psychology, perception, and diagnostic reasoning. Moreover, Plato's solution has always been accepted with many reservations. Polanyi's proposal is to use the concept of tacit knowledge:

> The *Meno* shows conclusively that if all knowledge is explicit, i.e, capable of being clearly stated, then we cannot know a problem or look for its solution. And the *Meno* also shows, therefore, that if problems nevertheless exist, and discoveries can be made by solving them, we can know things, and important things, that we cannot tell (Polanyi, 1966, p. 22).

May be Polanyi thinks the boy had a kind of precognition of the solution that he could not formulate in response to Socrates' direct query: "the kind of tacit knowledge that solves the paradox of the *Meno* consists in the intimation of something hidden, which we may yet discover" (Polanyi, 1966, p. 23).

Polanyi's epistemological allure is explained by an example derived from the history of science:

> The Copernicans must have meant to affirm a kind of foreknowledge when they passionately maintained during the hundred and forty years before Newton proved the point, that the heliocentric theory was not

---

quiry; nor again can he inquire about what he does not know, since he does not know about what he is to inquire" (Plato, 1977, 80e, p. 301).

> merely a convenient way of computing the paths of the planets, but was really true (Polanyi, 1966), p. 24).

The concept of tacit knowledge is also able to explain the fact that the scientist, when looking for "valid knowledge of a problem", a knowledge that is oriented toward constructing a "valid anticipation of the yet indeterminate implications of the discovery", will arrive at additional but "as yet undisclosed, perhaps as yet unthinkable, consequences" (Polanyi, 1966, p. 24).

The new meanings the *Meno*'s geometrical dialogue acquires anticipate further new models: the epistemological concept of "problem-solving" and tacit knowledge will be soon challenged again. Tacit knowledge is not so mysterious and non-analyzable, and can be modeled (the anthropologist has already met this problem). Polanyi's tacit knowledge may turn out to be explicit: the geometrical story of the *Meno* will acquire new scientific meanings in light of cognitive science and artificial intelligence (AI) (cf. the following section).

We have seen that *geometrical construction* plays a fundamental role in the *Meno* dialogue and subsequently in modern theories of tacit knowledge and problem-solving. It is important to note that it also plays an interesting role in modern philosophy. In the *Critique of Pure Reason* Kant's thought feeds on the reflections about geometrical construction and, at the same time, elaborates a new philosophical model of it. Hence the enigmas of geometrical construction are transformed by applying to them the new meanings of the great Kantian theory of imagination, *schematism* and synthetic *a priori*.

Let us look at some of its features. The problem of the construction of a concept is central to Kantian philosophy: when Kant has to study the problem of geometrical construction he measures himself with the same kind of problem that we have seen in the *Meno* dialogue. In the *Critique of Pure Reason,* Kant specifically studies the geometrical construction, which is just what Socrates acts out when drawing the square under the slave boy's eyes to provoke the suitable inferences which lead to the right solution. Kant says in the "Transcendental Doctrine of Method":

> To *construct* a concept means to exhibit *a priori* the intuition which corresponds to the concept. [...] Thus I construct a triangle, by representing the object which corresponds to this concept, either by imagination alone, in pure intuition, or in accordance therewith also on paper, in empirical intuition - in both cases completely *a priori*, without having borrowed the pattern from any experience (Kant, 1929, A713-B741, p. 577).

Hence, the possibility of drawing geometrical figures, like the *Meno* squares, is guaranteed by the activity *a priori* of imagination. More precisely, there is a universal schematic activity driven by the imagination, that

Kant sometimes classifies as a *rule*, sometimes as a *model*, and on other occasions as a *procedure*, which enables the passage from the pure geometrical concept to its sensible representation.

We can say that the activity of schematism is implicit, resulting from imagination and completely unknowable as regards to its ways of working, empty, and devoid of any possibility of being rationally analyzed. We may say, following Polanyi's line, that schematism is an activity of *tacit knowledge* (moreover, to this end Polanyi simply cites Kantian imagination). Kant says:

> This schematism of our understanding, in its application to appearances and their mere form, is an art concealed in the depths of the human soul, whose real modes of activity nature is hardly likely ever to allow us to discover, and to have open to our gaze (Kant, 1929, A141-B181, p. 183).

and then,

> [...] the result of the power of imagination [is] [...] a blind but indispensable function of the soul, without which we should have no knowledge whatsoever, but of which we are scarcely ever conscious (Kant, 1929, A78-B103, p.112).

It is well-known that Kant considers geometry as being constituted of synthetic *a priori* judgments; construction, indeed, is to "pass beyond":

> For I must not restrict my attention to what I am actually thinking in my concept of a triangle (this is nothing more than the mere definition); I must pass beyond it to properties which are not contained in this concept, but yet belong to it (Kant, 1929, A718-B746, p. 580).

There is no longer room for the doctrine of reminiscence: when I construct a square and draw its figure as much as I like, as Socrates does, I automatically "pass beyond" the pure concept to discover properties which I did not find before in the concept itself but which I then immediately verify belong to it. This process is guaranteed by the philosophical level of imagination and its schematic activity, which conditions the possibility of the geometrical construction itself[5]. Exploring the secrets of geometry and analyzing the cognitive virtues of geometrical construction, Kant comes to synthetic *a priori* judgments and to the concept of schematism, which are fundamental elements of transcendental philosophy itself. As for Plato, geometry acquires new extra-geometrical meanings: it is submitted to a philosophical translation, its meanings change and live again in a great theoretical project. Kant emphasizes that in mathematics, analytic and discursive methods are not able to produce new knowledge, since they are based upon dividing

[5] On Kant and geometry see Magnani, 2000.

concepts without going beyond them. Geometrical construction is shown to be the selection of all possible definitions given in a geometrically pure concept and that, among the chances offered by mere logical possibility, we have to choose the ones that are valid as regards the construction in the intuition, that is, with regard to a possible real application. The concept of construction refers in a broad sense to the general design of the whole "Transcendental Analytic", and subsequently to the design of transcendental philosophy.

2. GENERATE AND TEST

Tacit knowledge and Kantian schematism are not so mysterious and non-analyzable, and can be modeled. The important point is to work on the procedures and heuristics for solving problems and to provide rational and explicit frameworks, which are for instance applied to generating computational systems. Polanyi's tacit knowledge may turn out to be explicit: I will illustrate how the geometrical story of the *Meno* acquires new scientific meanings in light of cognitive science and artificial intelligence, so joining Peirce's landmark definition of abduction, modeled as a process of *generating and testing* (cf. chapter 2, section 1, this book). Moreover, this new perspective can introduce the role of inconsistencies in the geometrical problem solving process[6].

Polanyi's thesis, according to which we can know more than we can tell, is refuted by Simon in "Discussion: the Meno Paradox" (Simon, 1976, also in Simon, 1977), which clarifies a previous criticism of Polanyi's considerations on the *Meno* made by Bradie (Bradie, 1974). Let us remind ourselves of Polanyi's and *Meno*'s dilemma:

> [...] to search for the solution of a problem is an absurdity; for either you know what you are looking for, and then there is no problem; or you do not know what you are looking for, and then you cannot expect to find anything (Polanyi, 1966, p. 22).

Bradie thinks that one horn of this dilemma, that if you know what you are looking for, then there is no problem, is false. He gives the example of a researcher who is searching to refute Goldbach's conjecture: "He knows what he is looking for, namely an even number which is not the sum of two primes, but he still has the problem of finding one" (Bradie, 1974, p. 203).

Studying this example Simon is able to conclude that it is possible to model the problem-solving process in a completely explicit and formal way

[6] For an account on inconsistencies and anomalies in abductive reasoning cf. chapter 6, this book.

(Simon, 1977, p. 340). Simon provides a computational solution of the paradox in modern problem-solving terms: "our ability to know what we are looking for does *not* depend upon our having an effective procedure for finding it: we need only an effective procedure for *testing* candidates" (Simon, 1977, p. 339). If it is possible to have an effective procedure for testing, and a procedure for generating candidates, we will have a "problem", i.e. an unsolved problem, where we nevertheless "know what we are looking for" without actually possessing it. As Turner states, "In the case of Goldbach's conjecture, we can set up the following procedures: generate even numbers, generate numbers named by their prime factors, and make judgments of equality. The problem then can be defined as follows: 'find a number *k* generated by the first procedure that does not belong to the numbers generated by the second procedure'. Thus the example fits the 'general scheme for defining problem solutions prior to finding them'" (Turner, 1989, p. 86). The narrative models generated by Plato's (philosophical) text are now substituted by the new - rational - ones, consisting of the objective heuristics and procedures efficaciously represented in a rational medium: a theory and/or a program.

Following Turner (Turner, 1989, pp. 86-87), we may develop some further considerations. Does the slave boy possess a procedure for testing the solutions to the questions given by Socrates? Maybe he does not. Socrates shows that "he was wrong by making him go through certain inferential steps which lead to an inconsistency with his original answer"(p. 86), for example when he demonstrated that if the square has twice the area the sides will also doubled. From a computational point of view we can affirm that the boy possesses some "subroutines that generated each inferential step, plus subroutines that matched the outcomes and recorded a failure to match" (*ibid.*), and hence realizes that the answers generated by them had led to mistakes (this is the case of the first two answers: 1) the sides have to be doubled; 2) the sides have to be tripled). One might call into question whether Socrates has in fact merely interrogated the boy:

> [...] the slave boy apparently could not have performed the routine producing the result on his own; Socrates, by putting the pre-existing subroutines together into a routine measuring the area of squares in fact taught him something, namely the higher-level routine. It is perhaps less clear that the boy possessed a procedure for matching answers in order to test candidate solutions; after all, he didn't recognize his original answer was wrong until Socrates showed that, using the boy's matching and calculating subroutines, one comes to a different numerical result. It is the inconsistency that the boy recognizes as an error. So one may say that Socrates represented the geometrical problem in a way that the boy's pre-existing "numerical inconsistency recognition subroutine" could be in-

> voked, which it was not by Socrates' original query. Thus Socrates has taught at least one thing to the boy in this part of the dialogue, namely a higher level routine for calculating area, and perhaps another, the technique for the representation of the problem of area as a problem of adding the squares within a square, which Socrates supplies and the boy instantly accepts, and which is necessary for producing the inconsistency the boy recognizes (*ibid*).

Therefore, the boy is able to recognize the inconsistency as an error and, consequently, the related hypothesis as *provisional*, defeasible. The higher level routine taught by Socrates renders the pre-existing calculating subroutines (and subroutines for visual comparison) utilizable. From the point of view of the above description we may know that Socrates' version is false: on the contrary, Socrates has taught the boy something, we can say that Socrates has "programmed" the boy or the boy has "programmed" himself when he recognizes as valid the technique of adding the squares within a square (and rejects the answers generated by the pre-existing subroutines). The above description is in line with Simon's view according to which it is possible to analyze the problem-solving process in terms of formal and rational models, that can be further represented using suitable computational methods and tools.

On the contrary, to save Polanyi's account, we may postulate that the boy had a kind of precognition of the correct method: "Socrates' suggestions match with this tacit precognition which enables him to 'recognize as valid' Socrates' method of arriving at the answer. But precognition might be modeled as well, for example in terms of preference of previously experienced patterns" (p. 87). In conclusion, if we avoid Polanyi's interpretation in terms of an "empty" tacit knowledge, we may develop many explicit models (philosophical, psychological/mental, computational):

> One might emulate the slave boy in innumerable other ways: by beginning with a program for areas of all kinds of surfaces and blocking execution of any but the programs relating to squares, even the rules of recognition could be constructed differently, for they depend on what learning process to start with. If we think like Socrates, we might assume that all knowledge of means of calculating areas of surfaces is already present in the boy, and thus treat recognition as recollection (plus modellable error, as in mental processes), and construct a recognition subroutine such that the primary test the computer would use in "recognizing" would be to match the suggested learning with previously blocked parts of the program, and, if they match, unblock them. [...] One might suppose, for example, that mathematical concepts are not primary, but based on ontogenetically prior concepts and experiences that may or may not be

shared with persons raised differently - that there may be different paths to, and therefore different underlying inferential frameworks, onto which for thinking about mathematical concept is grafted. Some may reason casually, as the slave boy could, while others might find it more natural to derive their geometric results from numerical relations and to infer by operations on numbers (Turner, 1989).

To express the situation in terms of mental models we may emphasize the doubt occurring in the slave boy when faced with the failure of his first hypothesis of the tripled sides, which Socrates demonstrated to be wrong, and consider it like an *impasse*, like a transitional moment necessary for restructuring the whole problem (it is noteworthy that a psychologist like Ohlsson exhibits, to illustrate this cognitive situation, a geometrical construction very similar to that in the *Meno* - Ohlsson, 1984a, 1984b). At the same time, the *impasse* represents the awareness of the falseness and then of the defeasibility of prospected solutions (hypotheses), and highlights the essential uncertainty of the underlying reasoning. This is a "psychological" model very far from the "philosophical" meaningfulness of Socratic "doubt".

Chapter 2

Theoretical Abduction

1. WHAT IS ABDUCTION?

1.1 Abduction and retroduction

Philosophers of science in the twentieth century have traditionally distinguished between the logic of discovery and the logic of justification. Most have concluded that no logic of discovery exists and, moreover, that a *rational* model of discovery is impossible. In short, scientific discovery is irrational and there is no reasoning to hypotheses. A new abstraction paradigm aimed at unifying the different perspectives and providing some design insights for future ones is proposed here: the aim of this book is to emphasize the significance of *abduction* in order to illustrate the problem-solving process and to propose a unified epistemological model of scientific discovery (this chapter and chapters 6 and 7), diagnostic reasoning (chapter 4), and other kinds of creative reasoning (chapter 5).

This section aims to introduce the distinction, not previously analyzed, between two kinds of *abduction*, *theoretical* and *manipulative*, in order to provide an integrated framework to explain some of the main aspects of both creative and *model-based reasoning* effects engendered by the practice of science. The distinction appears to be extremely convenient: after having illustrated the *sentential models* together with their limitations (section 2), creativity will be viewed as the result of the highest cases of theoretical abduction showing the role of the so-called *model-based abduction* (section 3) (cf. Figure 1). Moreover, in chapter 3, I will delineate the first features of

what I call *manipulative abduction* by showing how we can find methods of constructivity at the experimental stage, where the recent epistemological tradition has settled the most negative effects of theory-ladenness.

Abduction is becoming an increasingly popular term in many fields of AI, such as diagnosis (Console and Torasso, 1991; de Kleer, Mackworth, and Reiter, 1990, 1992; Goel, 1989; Pople, 1973; Poole, 1988, 1989, 1991; Reggia, Dana, and Pearl, 1983; Reggia and Nau, 1984; Reiter, 1987; - especially in the field of medical KBSs - Josephson, et al., 1986; Magnani, 1988, 1992; Ramoni, et al., 1992; Torasso and Console, 1989), planning (Eshgi, 1988), natural language processing and motivation analysis (Charniak, 1988, Hobbs, et al., 1993), and logic programming (Kakas, Kowalski, and Toni, 1992). General considerations on abduction in science and AI can also be found in Gooding, 1996; Josephson and Josephson, 1994; Kuipers, 1998; Thagard, 1988, 1992, Shrager and Langley, 1990; on the relationships with probability and diagnosis in Peng and Reggia, 1987a and b, with ordinary life and philosophy in Harman, 1965, with historical and temporal reasoning in Fann, 1970, with emotions in O'Rorke, 1995 and Ortony, 1992, with narratives in Oatley, 1996, with decision-making in law courts, Pennington and Hastie, 1988.

Let us consider the following interesting passage, from an article by Simon (1965) and dealing with the logic of normative theories:

> The problem-solving process is not a process of "deducing" one set of imperatives (the performance programme) from another set (the goals). Instead, it is a process of selective trial and error, using heuristic rules derived from previous experience, that is sometimes successful in *discovering* means that are more or less efficacious in attaining some end. If we want a name for it, we can appropriately use the name coined by Peirce and revived recently by Norwood Hanson (1958): it is a *retroductive* process. The nature of this process - which has been sketched roughly here - is the main subject of the theory of problem-solving in both its positive and normative versions (Simon, 1977, p. 151).

Simon states that discovering means that are more or less efficacious in attaining some end are performed by a *retroductive* process. He goes on to show that it is easy to obtain one set of imperatives from another set by processes of discovery or retroduction, and that the relation between the initial set and the derived set is not a relation of logical implication. I completely agree with Simon: retroduction (that is abduction, cf. below) is the main subject of the theory of problem-solving and developments in the fields of cognitive science and artificial intelligence have strengthened this conviction.

Hanson (1958, p. 54) is perfectly aware of the fact that an enormous range of explanations (and causes) exists for any event:

> There are as many causes of x as there are explanations of x. Consider how the cause of death might have been set out by a physician as "multiple hemorrhage", by the barrister as "negligence on the part of the driver", by a carriage-builder as "a defect in the brakeblock construction", by a civic planner as "the presence of tall shrubbery at that turning".

The word "retroduction" used by Simon is the Hansonian neopositivistic one replacing the Peircian classical word abduction. Following Hanson's point of view Peirce "regards an abductive inference (such as 'The observed position of Mars falls between a circle and an oval, so the orbit must be an ellipse') and a perceptual judgment (such as 'It is laevorotatory') as being opposite sides of the same coin". It is also well-known that Hanson relates abduction to the role of patterns in reasoning and to the Wittgensteinian "Seeing that" (Hanson , 1958, p. 86)[1].

As Fetzer has recently stressed, from a philosophical point of view the main modes of argumentation for reasoning from premises to conclusions are expressed by these three general kinds of reasoning: *deductive* (demonstrative, non ampliative, additive), *inductive* (non-demonstrative, ampliative, non additive), *fallacious* (neither, irrelevant, ambiguous). Abduction, which expresses likelihood in reasoning, is a typical form of fallacious inference: "it is a matter of utilizing the principle of maximum likelihood in order to formalize a pattern of reasoning known as 'inference to the best explanation'" (Fetzer, 1990, p. 103)[2]. These different kinds of reasoning will be illustrated in the following section.

1.2 ST-MODEL and the syllogistic framework

Many reasoning conclusions that do not proceed in a deductive manner are *abductive*. For instance, if we see a broken horizontal glass on the floor[3] we might explain this fact by postulating the effect of wind shortly before: this is not certainly a deductive consequence of the glass being broken (a cat may well have been responsible for it). Hence, theoretical abduction is the process of *inferring* certain facts and/or laws and hypotheses that render some sen-

[1] In section 3.2 I will introduce the concept of model-based abduction; in chapter 5 two kinds of model-based abduction will be analyzed: visual, related to the role of abduction in visual imagery and perception, and temporal.

[2] On the inference to the best explanation see also Harman, 1965, 1967, Thagard, 1987, and Lipton, 1991.

[3] This event constitutes in its turn an *anomaly* that needs to be solved/explained.

tences plausible, that *explain* or *discover* some (eventually new) phenomenon or observation; it is the process of reasoning in which explanatory hypotheses are formed and evaluated. Moreover, we have to remember that although explanatory hypotheses can be elementary, there are also cases of composite, multipart hypotheses[4].

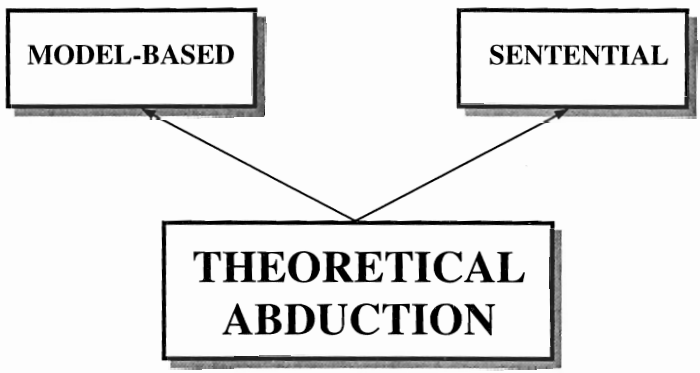

Figure 1. Theoretical abduction.

First, it is necessary to show the connections between abduction, induction, and deduction and to stress the significance of abduction to illustrate the problem-solving process. I think the example of diagnostic reasoning is an excellent way to introduce abduction. I have developed with others (Lanzola, et al., 1994; Ramoni, et al., 1992) an epistemological model of medical reasoning, called the *Select and Test Model* (ST-MODEL) (Magnani, 1992, Stefanelli and Ramoni, 1992) which can be described in terms of the classical notions of abduction, deduction and induction; it describes the different

[4] Anyway, some hypotheses are empty from the *explanatory* point of view: for example the generalization "every object in the population is female or male" does not explain that Maria is female, since it requires the additional knowledge that Maria is not male. The process of finding such generalizations has been called confirmatory (or descriptive) induction: "A typical form of explanatory induction is concept learning, where we want to learn a definition of a given concept *C* in terms of other concepts. This means that our inductive hypotheses are required to explain (logically entail) why particular individuals are *C*s, in terms of the properties they have. However, in the more general case of confirmatory induction we are not given a fixed concept to be learned. The aim is to learn relationships between any of the concepts, with no particular concept singled out. The formalization of confirmatory hypothesis formation cannot be based on logical entailment, as in Peirce's abduction. Rather, it is a quantitative form of degree of confirmation, which explains its name" (Flach and Kakas, 2000a).

roles played by such basic inference types in developing various kinds of medical reasoning (diagnosis, therapy planning, monitoring) but can be extended and regarded also as an illustration of scientific theory change. The model is consistent with the Peircian view about the various stages of scientific inquiry in terms of "hypothesis" generation, deduction (prediction), and induction.

The type of inference called abduction was also studied by Aristotelian syllogistics, as a form of ἀπαγωγή, and later on by mediaeval reworkers of syllogism. A hundred years ago, Peirce (*CP*, 1931-1958)[5] interpreted abduction essentially as an "inferential"[6] *creative process* of generating a new hypothesis. Abduction and induction, viewed together as processes of production and generation of new hypotheses, are sometimes called reduction, that is ἀπαγωγή[7]. As Lukasiewicz (1970) makes clear, "Reasoning which starts from reasons and looks for consequences is called *deduction*; that which starts from consequences and looks for reasons is called *reduction*". The celebrated example given by Peirce is Kepler's conclusion that the orbit of Mars must be an ellipse[8].

There are two main epistemological meanings of the word abduction (Magnani, 1988, 1991): 1) abduction that only generates "plausible" hypotheses (*selective* or *creative*) and 2) abduction considered as *inference to the best explanation*, which also evaluates hypotheses (see Figure 2)[9].

To illustrate from the field of medical knowledge, the discovery of a new disease and the manifestations it causes can be considered as the result of a creative abductive inference. Therefore, creative abduction deals with the whole field of the growth of scientific knowledge (Blois, 1984). This is irrelevant in medical diagnosis where instead the task is to *select* from an encyclopedia of pre-stored diagnostic entities (Ramoni, et al., 1992). We can call both inferences ampliative, selective and creative, because in both cases the reasoning involved amplifies, or goes beyond, the information incorporated in the premises.

All we can expect of our "selective" abduction, is that it tends to produce hypotheses for further examination that have some chance of turning out to be the best explanation. Selective abduction will always produce hypotheses that give at least a partial explanation and therefore have a small amount of initial plausibility. In the syllogistic view (see below) concerning abduction

[5] Cf. Frankfurt, 1958, Reilly, 1970, Fann, 1970, Davis, 1972, Ayim, 1974, Anderson, 1986, 1987, Kapitan, 1990, Hookway, 1992, Debrok, 1997, Roesler, 1997, Wirth, 1997.

[6] On the special meaning of the adjective "inferential" used by Peirce see section 3.2 below.

[7] Sometimes ἀπαγωγή is translated with retroduction, so it is simply referred to abduction (see above in this section).

[8] A clear reconstruction of Kepler's discovery is given in Gorman (1998).

[9] Further explanations of this bipolar distinction (and about the use herein of the concept of plausibility) are given below in this section.

as inference to the best explanation advocated by Peirce one might require that the final chosen explanation be the most "plausible" (cf. below, section 1.3).

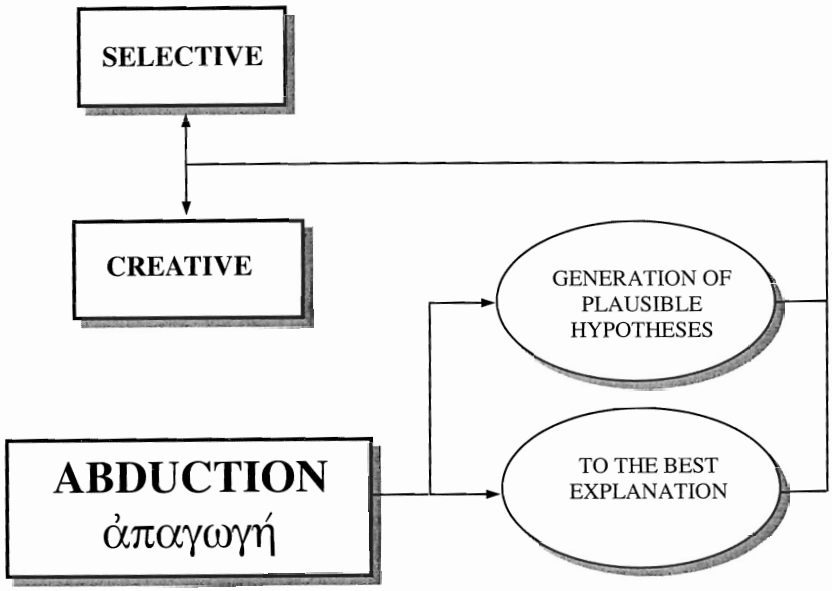

Figure 2. Creative and selective abduction.

Since the time of John Stuart Mill (1843), the name given to all kinds of non deductive reasoning has been induction, considered as an aggregate of many methods for discovering causal relationships. Consequently *induction* in its widest sense is an ampliative process of the generalization of knowledge. Peirce (1955a) distinguished various types of induction: a common feature of all kinds of induction is the ability to compare individual statements: using induction it is possible to synthesize individual statements into general laws - inductive generalizations - in a defeasible way, but it is also possible to confirm or discount hypotheses. Following Peirce, I am clearly referring here to the latter type of induction, that in the ST-MODEL is used as the process of reducing the uncertainty of established hypotheses by comparing their consequences with observed facts[10]. This perspective on hy-

[10] It is possible to treat every good inductive generalization as an instance of abduction (Josephson, 1994a, see also section 2.1 below). Some authors stress that abduction and induction derive from a common source, the hypothetical or non-deductive reasoning, others

pothesis testing in terms of induction is also known in philosophy of science as the "hypothetico-deductive method" (Hempel, 1966) (cf. p. 39, below) and is related to the idea of confirmation of scientific hypotheses, predominant in nepositivistic philosophy but also present in the falsificationist tradition (Popper, 1959).

Deduction is an inference that refers to a logical implication. Deduction may be distinguished from abduction and induction on the grounds that only in deduction is the truth of the conclusion of the inference guaranteed by the truth of the premises on which it is based. Deduction refers to the so-called non-defeasible arguments. It should be clear that, on the contrary, when we say that the premises of an argument provide partial support for the conclusion, we mean that if the premises were true, they would give us good reasons - but not conclusive reasons - to accept the conclusion. That is to say, although the premises, if true, provide some evidence to support the conclusion, the conclusion may still be false (arguments of this type are called inductive, or abductive, arguments).

All these distinctions need to be exemplified. To describe how the three inferences operate, it is useful to start with a very simple example dealing with diagnostic reasoning and illustrated (as Peirce initially did[11]), in *syllogistic terms* (see also Lycan, 1988):

1. If a patient is affected by a pneumonia, his/her level of white blood cells is increased.
2. John is affected by a pneumonia.
3. John's level of white blood cells is increased[12].

---

emphasize the various aspects that distinguish them, that is how specifically abduction and induction extend our knowledge. In other cases it is affirmed that all non-deductive reasoning is of the same type, which is called induction (Flach and Kakas, 2000a). Further classifications of inductive arguments have been proposed, such as arguments based on samples, (that is inductive generalizations), arguments from analogy, and statistical syllogisms (Salmon, 1990). Finally we have to remember that in the case of the so-called *inductive logic* (Carnap, 1950) the aim is to solve the problem of knowing the degree of belief we should attribute to the hypothetical conclusion H, given evidence E collected in the premises of an inductive argument, that is identified with the conditional probability $P(H|E)$. This formalization of the inductive support is also called *confirmation theory*: it does not deal with the problem of individuating the ways of "generating" inductive hypotheses but refers to a logic of hypothesis "evaluation".

[11] Some authors (Flach and Kakas, 2000a, Aliseda, 2000) distinguish between Peircian early syllogistic theory and his later inferential one, in which abduction refers to the hypothesis formation component of explanatory reasoning (see section 3 below).

[12] The famous syllogistic example given by Peirce is:

1. All beans from this bag are white.
2. These beans are from this bag.
3. These beans are white.

(This syllogism is known as Barbara).

By deduction we can infer (3) from (1) and (2). Two other syllogisms can be obtained from Barbara if we exchange the conclusion (or Result, in Peircian terms) with either the major premise (the Rule) or the minor premise (the Case): by induction we can go from a finite set of facts, like (2) and (3), to a universally quantified generalization - also called categorical inductive generalization, like the piece of hematologic knowledge represented by (1)[13]. Starting from knowing – selecting – (1) and "observing" (3) we can infer (2) by performing a selective abduction[14]. The abductive inference rule corresponds to the well-known fallacy called affirming the consequent (simplified to the propositional case)

$$\frac{\varphi \rightarrow \psi \quad \psi}{\varphi}$$

It is useful to give another example, describing an inference very similar to the previous one:

1. If a patient is affected by a beta-thalassemia, his/her level of hemoglobin A2 is increased.
2. John is affected by a beta-thalassemia.
3. John's level of hemoglobin A2 is increased.

Such an inference is valid, that is not affected by uncertainty, since the manifestation (3) is pathognomonic for beta-thalassemia (as expressed by the biconditional in $\varphi \leftrightarrow \psi$). This is a special case, where there is not abduction because there is not "selection", in general clinicians very often have to deal with manifestations which can be explained by different diagnostic hypotheses: in these case the inference rule corresponds to

$$\frac{\varphi \leftrightarrow \psi \quad \psi}{\varphi}$$

[13] We can consider this inference a sort of generalization from a sample of patients [or of beans] to the whole population of them [or of beans in the bag].

[14] We have to remark that at the level of the syllogistic treatment of the subject Peirce calls this kind of argumentation "hypothesis"; he will introduce the term abduction only in his later theory.

Thus, *selective abduction* is the making of a preliminary guess that introduces a set of plausible diagnostic hypotheses, followed by deduction to explore their consequences, and by induction to test them with available patient data, (1) to increase the likelihood of a hypothesis by noting evidence explained by that one, rather than by competing hypotheses, or (2) to refute all but one (cf. Figure 3).

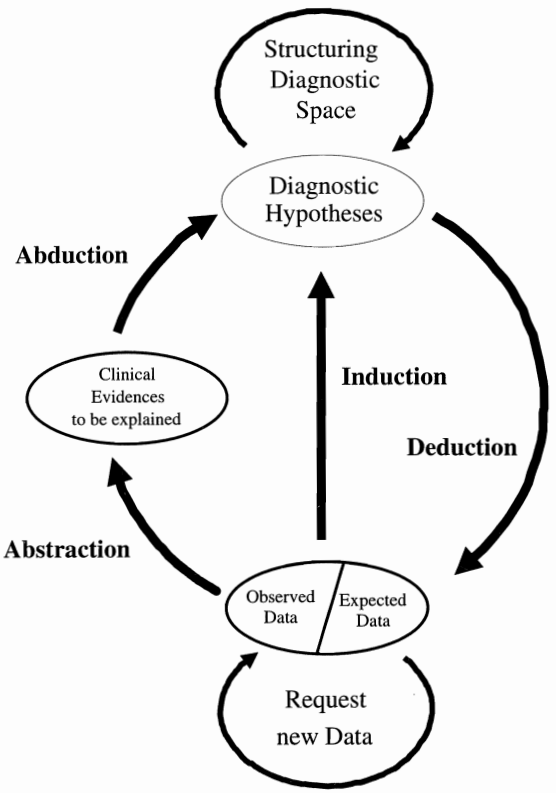

Figure 3. The epistemological model of diagnostic reasoning.

If during this first cycle new information emerges, hypotheses not previously considered can be suggested and a new cycle takes place. In this case the *nonmonotonic* character of abductive reasoning is clear and arises from the logical unsoundness of the inference rule: it draws defeasible conclusions from incomplete information[15]. All recent logical accounts ("deductive")

[15] A logical system is monotonic if the function *Theo* that relates every set of wffs to the set of their theorems holds the following property: for every set of premises *S* and for every set

concerning abduction have pointed out that it is a form of nonmonotonic reasoning. It is important to allow the guessing of explanations for a situation, in order to discount and abandon old hypotheses, so as to enable the tentative adoption of new ones, when new information about the situation makes them no longer the best.

As Stephanou and Sage (1987) pointed out, uncertainty and imperfect information are fundamental characteristics of the knowledge relative to hypothetical reasoning. The nonmonotonic character of the ST-MODEL arises not only from the above mentioned nonmonotonic character of deductive inference type involved in it, but also from the logical unsoundness of the ascending part of the cycle guessing hypotheses to be tested. Doyle (1988) pointed out that, since their unsoundness, these guesses do not exhibit the truth-preservative behavior of ideal rationality characterizing the incremental deduction of classical logic, but the nonmonotonic behavior of limited rationality of commonsense reasoning (Simon, 1969), that allows to discharge and abandon old hypotheses to make possible the tentative adoption of new ones. Notice that this adoption is not merely tentative but rationally tentative, in the sense that, just as abduction, it is based on a reasoned selection of knowledge (Truesdell, 1984) and on some preference criteria which avoid the combinatorial explosion of hypotheses generation.

One of the principal means of limiting rationality is indeed to limit efforts by directing attention to some areas and ignoring others. This character matches exactly with the ability of an expert in generating a small set of hypotheses to be carefully tested. But in such a case, the expert has to be ready to withdraw paths of reasoning when they diverge from the correct path, that is from the path that would have taken the expert had considering the ignored knowledge portions. In such a way, the nonmonotonic character turns out as a foundational epistemological feature of the ST-MODEL of medical reasoning, since this nonmonotonic character is the result not of a mere lack of information but of a reasoned limiting of information imposed by its own logical unsoundness.

Modern logic allows us to account for this dynamic behavior of abduction by the concept of *belief revision*. Belief revision (Alchourrón, Gärdenfors, and Makinson, 1985) is a dynamic notion dealing with the current stage of reasoning. At each stage of reasoning, if it is correct, a belief is held on

---

of premises S', $S \subseteq S'$ implies $Teo(S) \subseteq (S')$. Traditional deductive logics are always monotonic: intuitively, adding new premises (axioms) will never invalidate old conclusions. In a nonmonotonic system, when axioms, or premises, increase, their theorems do not (cf. Ginsberg, 1987; Lukaszewicz, 1990; Magnani and Gennari, 1997). Following this deductive nonmonotonic view of abduction, we can stress the fact that in actual abductive medical reasoning, when we increase symptoms and patients' data [premises], we are compelled to abandon previously derived plausible diagnostic hypotheses [theorems], as already - epistemologically - illustrated by the ST-MODEL.

the basis that that reasoning is justified, even if subsequent stages dictate its retraction. A logic of belief for abduction has been proposed by Levesque (Levesque, 1989), and the role of belief revision functions in abduction has already been studied by Jackson (Jackson, 1990) (see section 2, below). Clearly abduction in medical diagnostic reasoning is an example of non-monotonic deduction.

1.3 Abduction as hypothesis generation, abduction as hypothesis generation and evaluation

As stated above, there are two main epistemological meanings of the word abduction: (1) abduction that only generates plausible hypotheses (*selective* or *creative*) – this is the meaning of abduction accepted in my epistemological model – and (2) abduction considered as *inference to the best explanation*, that also evaluates hypotheses. In the latter sense the classical meaning of selective abduction as inference to the best explanation (for instance in medicine, to the best diagnosis) is described by the complete abduction–deduction–induction cycle. This distinction needs further clarification.

It is clear that the two meanings are related to the distinction between hypothesis generation and hypothesis evaluation, so abduction is the process of generating explanatory hypotheses, and induction matches the hypothetico-deductive method of hypothesis testing (1st meaning). However, we have to remember (as we have already stressed) that sometimes in the literature (and also in Peirce's texts) the word abduction is also referred to the whole cycle, that is as an inference to the best explanation (2nd meaning).

As Thagard has pointed out (Thagard, 1988, p. 53) the question was controversial in Peirce's writings too. Before the 1890s, Peirce discussed the hypothesis as follows: "Hypothesis is where we find some very curious circumstance which would be explained by the supposition that it was the case of a certain general rule, and thereupon adopt that supposition" (*CP*, 2. 624). When Peirce replaced hypothesis with abduction he said that it "furnishes the reasoner with the problematic theory which induction verifies" (*CP*, 2. 776). Thagard ascribes to the editors of Peirce's work the responsibility for having clouded this change in his thinking by including discussions of hypothesis under the heading of "Abduction", "obscuring his shift from the belief that inference to an explanatory hypothesis can be a kind of justification to the weaker view that it is only a form of discovery" (Thagard, 1988, p. 53). The need for a methodological criterion of justification is caused by the fact that an abduced hypothesis that explains a certain puzzling fact should not be accepted because of the possibility of other explanations.

Having a hypothesis that explains a certain number of facts is far from a guarantee of being true.

It could be said that maintaining this rigid distinction between generation and evaluation phase could lead to the generation of many uninteresting and useless hypotheses. It is important to note that already at the generation phase many evaluation considerations can be present and intertwined, so that we can say (as we said above in the previous section) that abduction considered as a way of generating hypotheses is immediately a generation of "plausible" hypotheses. I think this controversial status of abduction is related to a confusion between the epistemological and cognitive levels, and to a lack of explanation as to why people sometimes deviate from normative epistemological principles. An analysis of the differences between epistemological and cognitive levels would help to clarify the issue. By analyzing the case of real physicians reasoning performances, together with the structure of diagnostic medical knowledge-based systems (chapter 4, section 2), we will see that the rigid distinction between the generation and evaluation phase is more suitable to illustrate (and build) artificial reasoning systems than to account for real human thinking (where the two phases are usually highly integrated).

The combinatorial explosion of alternatives that has to be considered makes the task of finding the best explanation very costly. Peirce surely thinks abduction has to be *explanatory*, but also capable of experimental *verification* (that is evaluated inductively, cf. the model above), and *economic* (this includes the cost of verifying the hypothesis, its basic value, and other factors). Consequently, to achieve the best explanation, it is necessary to have or establish a set of criteria for evaluating the competing explanatory hypotheses reached by creative or selective abduction. Evaluation has a multi-dimensional and comparative character.

Consilience (Thagard, 1988) can measure how much a hypothesis explains, so it can be used to determine whether one hypothesis explains more of the evidence (for instance, empirical or patient data) than another: thus, it deals with a form of corroboration. In this way a hypothesis is considered more consilient than another if it explains more "important" (as opposed to "trivial") data than the others do. In inferring the best explanation, the aim is not the sheer amount of data explained, but its relative significance. The assessment of relative importance presupposes that an inquirer has a rich background knowledge about the kinds of criteria that concern the data. The evaluation is strongly influenced by Ockham's razor: *simplicity* too can be highly relevant when discriminating between competing explanatory hypotheses; for example, it deals with the problem of the level of conceptual complexity of hypotheses when their consiliences are equal.

Explanatory criteria are needed because the rejection of a hypothesis requires demonstrating that a competing hypothesis provides a better explanation. Clearly, in some cases - for instance when choosing scientific hypotheses or theories, where the role of "explanation" is dominant - conclusions are reached according to rational criteria such as consilience or simplicity. We will see in chapter 4 (section 1), that in the case of selecting diagnostic hypotheses the epistemic reasons are dominant, whereas, in the case of selecting therapies, epistemic reasons are of course intertwined with pragmatic and ethical reasons, which will play a very important role. Hence, in reasoning to the best explanation, motivational, ethical or pragmatic criteria cannot be neglected. Indeed the context suggests that they are unavoidable: as we have just mentioned, this is for example true in some part of medical reasoning (in therapy planning), but scientists that must discriminate between competing scientific hypotheses or competing scientific theories have to recognize that sometimes they too are conditioned by motivationally biasing their inferences to the best explanation. Some epistemologists, like Kuhn (1970) and Feyerabend (1993), argued that in science these extra-rational motivation are unavoidable.

For example, the so-called theory of *explanatory coherence* (Thagard, 1989, 1992) introduces seven principles able to grasp the type of plausibility that occurs in the acceptation of new hypotheses and theories in science (but also, with slight modifications, in many other fields like conceptual combination; adversarial problem-solving, when one has to infer an opponent's intentions; analogical reasoning; jury decisions in murder trials: contemporary debates about why the dinosaurs became extinct; psychological experiments on how beginning students learn physics; ethical deliberation; emotional decision); the theory is susceptible to be treated at the computational level using a local connectionist network.

Josephson has stressed that evaluation in abductive reasoning has to be referred to the following criteria

1. How a hypothesis surpasses the alternatives.
2. How the hypothesis is good in itself.
3. Its confidence in the accuracy of the data.
4. How thorough was the search for alternative explanations (Josephson, 2000b).

There is no agreement about which preference criteria to adopt. Hendricks and Faye (1999), speak, in the case of science, about correctness (concerning the world that it is investigating), empirical adequacy, simplicity (different kinds of), unification, consistency, practical usability, economy. Poole and Rowen (1990) list several criteria that have been proposed in the

literature and it can be shown that some of these preference criteria are conflicting, i.e. in the same situation, they favor different conjectures. The problem is that all the proposed criteria do not work in all situations: they are in some sense context dependent. For instance, the (syntactic) criterion of minimality described by the sentential models of abduction (cf. the following section), is useless when the conjecture at hand is (syntactically) as simple as the conflicting conjectures.

We can also use mathematical probability to select among hypotheses evaluating them (Bayes's Theorem itself can be viewed as a modality for weighing alternative hypotheses, Krauß, Martignon, and Hoffrage, 1999, of course in case the appropriate knowledge of probabilities is present).

The epistemological model (ST-MODEL) previously illustrated should also be regarded as a very simple and schematic illustration of scientific theory change. In this case selective abduction is replaced by creative abduction and there exists a set of competing theories instead of diagnostic hypotheses. Furthermore the language of background scientific knowledge should be regarded as open: in the case of competing theories, as they are studied using the epistemology of theory change, we cannot - contrary to Popper's initial viewpoint (Popper, 1959) - reject a theory simply because it fails occasionally. If for example such a theory is simpler and explains more significant data than its competitors, then it can be accepted as the best explanation.

As already stressed, in accordance with the epistemological model previously illustrated, medical reasoning may be broken down into two different phases: first, patient data is abstracted and used *to select hypotheses*, that is hypothetical solutions of the patient's problem (selective abduction phase); second, these hypotheses provide the starting conditions for forecasts of expected consequences which should be compared to the patient's data in order *to evaluate* (corroborate or eliminate) those hypotheses which they come from (deduction-induction cycle).

If we consider the epistemological model as an illustration of medical diagnostic reasoning, the modus tollens is very efficacious because of the fixedness of language that expresses the background medical knowledge: a hypothesis that fails can nearly always be rejected immediately.

When Buchanan illustrates the old epistemological method of induction by elimination (and its computational meaning, evident if we add a "heuristic search"[16], to limit the exhaustive enumeration of the derived hypotheses), first advanced by Bacon and Hooke and developed later on by J. Stuart Mill, he is referring implicitly to my epistemological framework in terms of abduction, deduction and induction, as illustrative of medical diagnostic reasoning:

[16] Further clarifications are given in chapter 4, section 2.

> The method of systematic exploration is [...] very like the old method of induction by elimination. Solutions to problems can be found and proved correct, in this view, by enumerating possible solutions and refuting all but one. Obviously the method is used frequently in contemporary science and medicine, and is as powerful as the generator of possibilities. According to Laudan, however, the method of proof by eliminative induction, advanced by Bacon and Hooke, was dropped after Condillac, Newton, and LeSage argued successfully that it is impossible to enumerate exhaustively all the hypotheses that could conceivably explain a set of events. The force of the refutation lies in the open-endedness of the language of science. Within a fixed language the method reduces to modus tollens [...]. The computational method known as heuristic search is in some sense a revival of those old ideas of induction by elimination, but with machine methods of generation and search substituted for exhaustive enumeration. Instead of enumerating all sentences in the language of science and trying each one in turn, a computer program can use heuristics enabling it to discard large classes of hypotheses and search only a small number of remaining possibilities (Buchanan, 1985, pp. 97-98).

2. THE SENTENTIAL FRAMEWORK

Sentential abduction can be rendered in different ways. For example, in the syllogistic framework we have just described abduction is considered like something propositional and as a type of fallacious reasoning. If we want to model abduction in a computational logic-based system, the fundamental operation is search (Thagard, 1996). When there is a problem to solve, we usually face several possibilities (hypotheses) and we have to select the suitable one (cf. selective abduction, above). Accomplishing the assigned task requires that we have to search through the space of possible solutions to find the desired one. In this situation we have to rely on heuristics, that are rules of thumb expressed in sentential terms that help in arriving at satisfactory choices without considering all the possibilities. An example of simple heuristic could be a rule like "Wear green socks with white pants but not with blue pants". The famous concept of *heuristic search*, which is at the basis of many computational systems based on propositional rules (cf. section 5, this chapter) can perform this kind of sentential abduction (selective). Of course other computational tools can be used to this aim, like neural and probabilistic networks, and frames-like representations.

Another important way of modeling abduction in a sentential way resorts to the development of suitable logical systems, that in turn are computationally exploitable in the area of the so-called logic programming (cf. section 2.1, be-

low). Many attempts have been made to model abduction by developing some formal tools in order to illustrate its computational properties and the relationships with the different forms of deductive reasoning (see, for example, Bylander, et al., 1991; Console, Theseider Dupré, and Torasso, 1991; Coz and Pietrsykowski, 1986; De Raedt and Bruynooghe, 1991; Jackson, 1989; Kakas, Kowalski, and Toni, 1992; Konolige, 1992; Josephson, 1994a; Levesque, 1989; O'Rorke, 1994; Poole, 1988, 1989; Reiter, 1987; Shanahan, 1989; Reiter and De Kleer, 1987). Some of the more recent formal models of abductive reasoning, for instance Boutilier and Becher (1995), are based on the theory of the epistemic state of an agent (Alchourrón, et al., 1985; Gärdenfors, 1988, 1992), where the epistemic state of an individual is modeled as a consistent set of beliefs that can change by expansion and contraction (*belief revision framework*)[17]. We shall discuss the nature of the kinds of inconsistencies captured by these formalisms and show how they do not adequately account for some roles played by anomalies, conflicts, and contradictions in many forms of explanatory creative reasoning (cf. chapter 6).

Deductive models of abduction may be characterized as follows (Figure 4). An explanation for β relative to background theory T will be any α that, together with T, entails β (normally with the additional condition that $\alpha \cup T$ be consistent). Such theories are usually generalized in many directions: first of all by showing that explanations entail their conclusions only in a *defeasible* way (there are many potential explanations), so joining the whole area of the so-called nonmonotonic logic or of the probabilistic treatments; second, trying to show how some of the explanations are relatively implausible, elaborating suitable technical tools (for example in terms of modal logic) able to capture the notion of preference among explanations.

Hence, we may require that an explanation makes the observation simply sufficiently probable (Pearl, 1988) or that the explanations that are more likely will be the "preferred" explanations: the involvement of a cat in breaking the glass is less probable than the effect of wind. Finally, the deductive model of abduction does not authorize us to explain facts that are inconsistent with the background theory notwithstanding the fact that these explanations are very important and ubiquitous, for instance in diagnostic applications, where the facts to be explained contradict the expectation that the system involved is working according to specification.

Boutilier and Becher (1995) provide a formal account of the whole question in term of belief revision: if believing A is sufficient to induce belief in B, then A (epistemically) *explains* B; the situation can be semantically illustrated in terms of an ordering of plausibility or normality which is able to represent the epistemic state of an agent. The conflicting observations will

[17] Levi's theory of suppositional reasoning is also related to the problem of "belief change" (Levi, 1991, 1996).

require explanations that compel the agent to withdraw its beliefs (hypotheses), and the derived conditional logic is able to account for explanations of facts that conflict with the existing beliefs. The authors are able to reconstruct, within their framework, the two main paradigms of model-based diagnosis, abductive (Poole, 1988, 1991), and consistency-based (de Kleer, et al., 1990; Reiter, 1987), providing an alternative semantics for both in terms of a plausibility ordering over possible worlds.

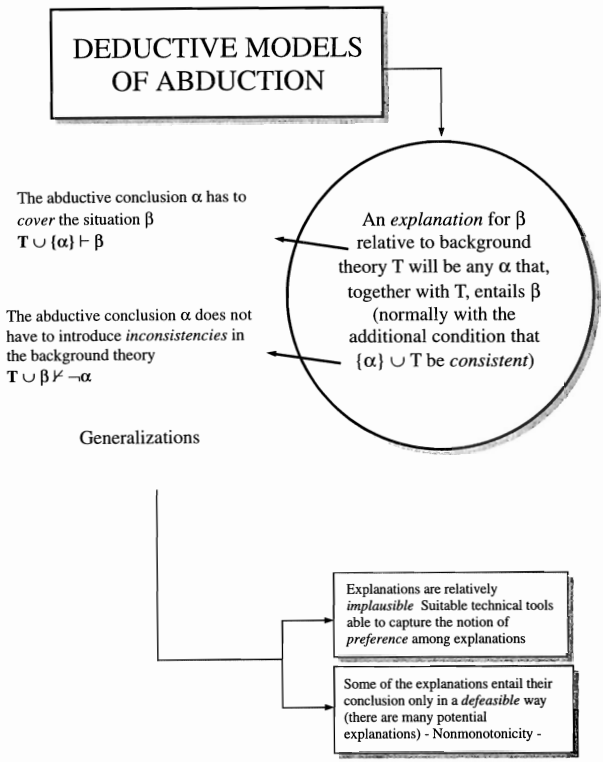

Figure 4. Deductive models of abductive reasoning.

Let us resume the kinds of change considered in the original belief revision framework (see Figure 5). The *expansion* of a set of beliefs K taken from some underlying language (considered to be the closure of some finite set of premise KB, or *knowledge base*, so $K = Cn(KB)$) by a piece of new information A is the belief set $K + A = Cn(K \cup AB)$. The addition happens "regardless" of whether the larger set is *consistent*. The case of *revision* happens when $K \models \neg A$, that is when the new A is *inconsistent* with K and we

want to maintain consistency: some beliefs in K must be withdrawn before A can be accommodated: $K - A$. The problem is that it is difficult to detect which part of K has to be withdrawn. The least "entrenched" beliefs in K should be withdrawn and A added to the "contracted" set of beliefs. The loss of information has to be as small as possible so that "no belief is given up unnecessarily" (Gärdenfors, 1988). Hence, *inconsistency resolution* in belief revision framework is captured by the concept of revision. Another way of belief change is the process of *contraction*. When a belief set K is contracted by A, the resulting belief set $K + A$ is such that A is no longer held, without adding any new fact[18].

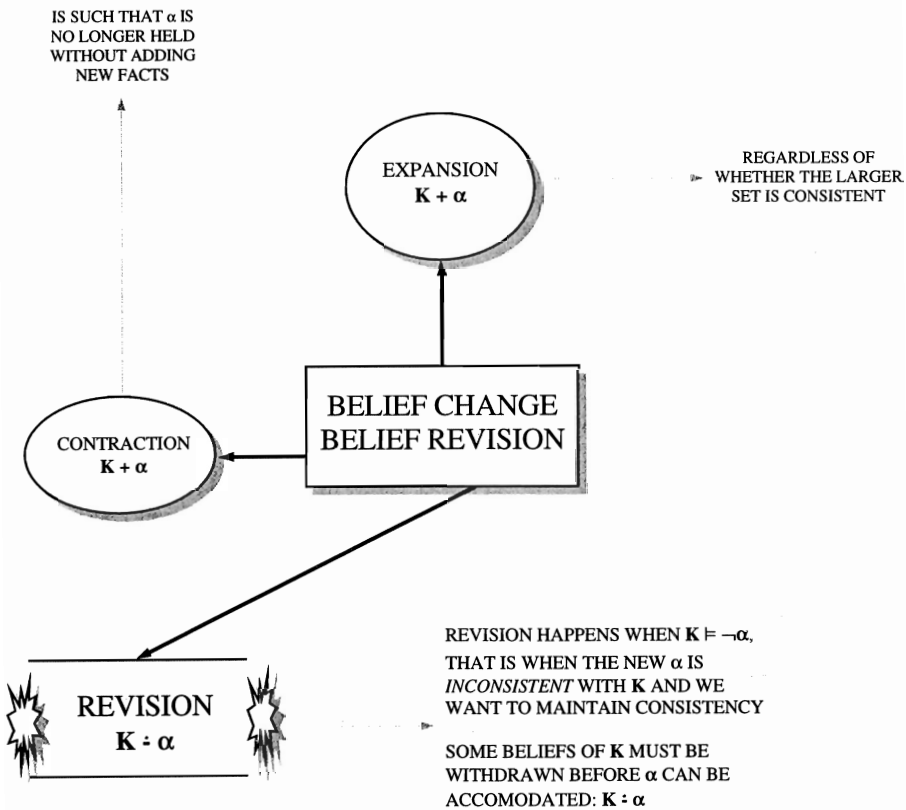

Figure 5. Belief revision.

[18] Aliseda (1997, 2000) makes use of the belief revision framework to construct a theory of the epistemic transition between the states of doubt and belief able to account for many aspects of abductive reasoning. On the relationships between belief revision dynamics in data bases and abduction cf. Aravindan and Dung, 1995.

After having explained the distinction between predictive explanations and "might" explanations, that merely allow an observation, and do not predict it, Boutilier and Becher show in the cited article how model-based diagnoses can be accounted for in terms of their new formal model of belief revision.

The *abductive* model-based reasoning[19] (Poole, 1988, 1991, Brewka, 1989) illustrated by some models, such as Poole's Theorist, allows many possible explanations, weak and predictive (so presenting a paraconsistent behavior: a non-predictive hypothesis can explain both a proposition and its negation). This old model, embedded in the new formal framework, acquires the possibility of discriminating certain explanations as preferred to others.

Reiter's *consistency-based* diagnosis (1987) is devoted to ascertain why a correctly designed system is not working according to its features. Because certain components may fail, the system description also contains some abnormality predicates (the absence of them will render the description inconsistent with an observation of an incorrect behavior). The consistency-based diagnosis concerns any set of components whose abnormality makes the observation consistent with the description of the system. A principle of parsimony is also introduced to capture the idea of preferred explanations/diagnoses. Since the presence of fault models renders Reiter's framework incorrect, new more complicated notions are introduced in (de Kleer, et al., 1990), where the presence of a complete fault model ensures that predictive explanations may be given for "every'' abnormal observation. Without any description of correct behavior any observation is consistent with the assumption that the system works correctly. Hence, a complete model of correct behavior is necessary if we want the consistency-based diagnosis to be useful. The idea of consistency that underlies some of the more recent deductive consistency-based models of selective abduction (diagnostic reasoning) is the following: any inconsistency (anomalous observation) refers to an aberrant behavior that can usually be accounted for by finding some set of components of a system that, if behaving abnormally, will entail or justify the actual observation. The observation is anomalous because it contradicts the expectation that the system involved is working according to specification. This types of deductive models go beyond the mere treatment of selective abduction in terms of preferred explanations and include the role of those components whose abnormality makes the observation (no longer anomalous) consistent with the description of the system (Boutilier and Becher, 1995; Magnani, 1999a).

Without doubt the solution given by Boutilier and Becher furnishes a more satisfying qualitative account of the choice among competing explana-

[19] Please distinguish here the technical use of the attribute *model-based* from the epistemological-cognitive one I introduced in the previous section.

tions than Gärdenfors' in terms of "epistemic entrenchment"[20] which tries to capture the idea of an ordering of beliefs according to our willingness to withdraw them when necessary. Moreover, the new formal account in terms of belief revision is very powerful in shedding new light on the old model-based accounts of diagnostic reasoning.

The framework of belief revision is sometimes called *coherence approach* (Doyle, 1992). In this approach, it is important that the agent holds some beliefs just as long as they are consistent with the agent's remaining beliefs. Inconsistent beliefs do not describe any world, and so are unproductive; moreover, the changes must be epistemologically conservative in the sense that the agent maintains as many of its beliefs as possible when it adjusts its beliefs to the new information. It is contrasted to the *foundations approach*, according to which beliefs change as the agent adopts or abandons satisfactory reasons (or justifications). This approach is exemplified by the well-known "reason maintenance systems" (RMS) or "truth maintenance systems" (TMS) (Doyle, 1979), elaborated in the area of artificial intelligence to cooperate with an external problem solver. In this approach, the role of inconsistencies is concentrated on the negations able to invalidate justifications of beliefs; moreover, as there are many similarities between reasoning with incomplete information and acting with inconsistent information, the operations of RMS concerning revision directly involve logical consistency, seeking to solve a conflict among beliefs. The operations of *dependency-directed backtracking* (DDB) are devoted to this aim: RMS informs DDB whenever a contradiction node (for instance a set of beliefs) becomes believed, then DDB attempts to remove reasons and premises, only to defeat nonmonotonic assumptions: "If the argument for the contradiction node does not depend on any of these (i.e., it consists entirely of monotonic reasons), DDB leaves the contradiction node in place as a continuing belief" (Doyle, 1992, p. 36), so leaving the conflicting beliefs intact if they do not depend on defeasible assumptions, and presenting a paraconsistent behavior.

Both in the coherence and foundations approach the changes of state have to be epistemologically conservative: as already said above the agent maintains as many of its beliefs as possible when it adjusts its beliefs to the new information, thus following Quine's idea of "minimun mutilation" (Quine, 1979). We have now to notice some limitations of the formal models in accounting for other kinds of inconsistencies embedded in many reasoning tasks.

This kind of sentential frameworks[21] exclusively deals with selective abduction (diagnostic reasoning) and relates to the idea of preserving *consis-*

[20] Which of course may change over time or with the state of belief.

[21] Sentential abduction is also active at the level of everyday natural language, where we generate creative (or simply new) narratives (see chapter 6, section 2.6).

tency. Exclusively considering the sentential view of abduction does not enable us to say much about creative processes in science. It mainly refers to the *selective* (diagnostic) and *explanatory* aspects of reasoning and to the idea that abduction is mainly an inference to the best explanation: when used to express the creativity events it is either empty or replicates the well-known *Gestalt* model of radical innovation. It is empty because the sentential view stops any attempt to analyze the creative processes.

Already in the Peircian syllogistic and sentential initial conception of abduction – as the fallacy of affirming the consequent, described in the previous section, we immediately see it is perfectly compatible with the *Gestalt* model of discovery. In both cases the event of creating something new (for example a new concept) is considered so radical and instantaneous that its irrationality is immediately involved. In this case the process is not considered as algorithmic: "the abductive suggestion comes to us like a flash. It is an act of insight, although of extremely fallible insight" (*CP* 5.181). Moreover, Peirce considers abduction as "a capacity of guessing right", and a "mysterious guessing power" common to all scientific research (*CP*, 6.530).

Notwithstanding its non-algorithmic character it is well known that for Peirce abduction is an *inferential process* (cf. below, section 3.2, for an explanation of the exact meaning of the word "inference" in Peircian thought): abduction

> is logical inference [...] having a perfectly definite logical form. [...] The form of inference, therefore, is this:
> The surprising fact, *C*, is observed;
> But if *A* were true, *C* would be a matter of course,
> Hence, there is reason to suspect that *A* is true (*CP* 5.188-189, 7.202).

C is true of the actual world and it is surprising, a kind of state of doubt we are not able to account for by using our available knowledge. *C* can be simply a *novel* phenomenon or may be in conflict with the background knowledge, that is *anomalous*[22].

The abductive inference includes all the operations whereby hypotheses and theories are constructed (*CP* 5.590) (see also Hintikka, 1998). Hence abduction has to be considered as a kind of *ampliative* inference that is not logical and truth preserving (in the sense of deductive): indeed valid deduction does not yield any new information, for example new hypotheses previously unknown (on the possible abductive sides of some forms of deduction see also section 4 below).

[22] The importance of surprise, novelty, and anomaly in abductive reasoning is analyzed in chapter 6.

To conclude, if we want to provide a suitable framework for analyzing the most interesting cases of conceptual changes in science we do not have to limit ourselves to the sentential view of theoretical abduction but we have to consider a broader *inferential* one which encompasses both sentential and what I call *model-based* sides of creative abduction (see, for details, section 3).

2.1 Abduction and induction in logic programming

The syllogistic account of abduction we described above is the starting point of much research in AI and logic programming devoted to perform tasks such as diagnosis in medical reasoning and planning. In these logical and computational accounts abduction and induction are considered as separate forms of reasoning related to different tasks. Consequently, the distinction is very variable, context-dependent and different from the one we have seen operating in Peircian texts (where, as illustrated in section 1.3, abduction is especially - but not only - viewed as hypothesis generation and induction as a logic of hypothesis evaluation).

In a recent book edited by Flach and Kakas (2000), many interesting contributions are dedicated to the analysis of the distinction between *abductive* and *inductive* reasoning[23]. Usually in these types of research abductive hypotheses are considered as providing explanations and inductive hypotheses as providing generalizations: this explains, for example, why diagnosis is generally considered in AI like abductive and concept learning from examples inductive[24]. Usually abduction is regarded as reasoning from specific observations to their explanations, and induction as a Millian enumerative induction from samples to general statements.

In the case of the *abductive logic programming* (ALP), and assuming a common first-order language, possible abductive hypotheses are built from specific non-observable predicates Δ called *abducibles* (suitably distinguished from observable predicates). The problem is to be able to "select" among the so-called abductive extensions $T(\Delta)$ of T in which the given observation to be explained holds, by selecting the corresponding formula Δ. On the contrary, in the case of *inductive logic programming* (ILP) the problem is to "select" a *generalizing* hypothesis able to entail additional observable information on unobserved individuals (that is predictions), finding new individuals for which the addition of the hypothesis to our knowledge is necessary to derive some observable properties for them (Console and Saitta,

[23] Cfr. also the clear article by Console, Theseider Dupré, Torasso, 1991.

[24] An overview on the relationships between abduction and induction is given in Bessant (2000). Cfr. also Abe; Christiansen; Inoue and Haneda; Mooney; Poole; Psillos; Sakama; Yamamoto, in the recent Flach and Kakas (2000).

2000). In the first case the abductive explanation Δ needs a given theory T, so it is "relative" to a "certain" theory T from which it has produced. In the case of induction the explanation does not depend on a particular theory: we can say that "all the beans from this bag are white" (Peirce's example, see above, section 1.2, this chapter), is an explanation for why the observed beans from the bag are white: this explanation does is in accordance with a particular model of the "world of beans" (Flach and Kakas, 2000a).

Moreover, we can say that abductions explain a phenomenon by indicating enabling conditions like causes (this explains the fact that abduction needs a domain theory, often a causal theory, while induction does not)[25]:

> If we want to explain, for instance, that the light appears in a bulb when we turn a switch on, an inductive explanation would say that this is because it happened hundreds of time before, whereas the abductive one can supply an explanation in terms of the electric current flowing into the bulb filament. If, at some moment, turning the switch on does not let the light bulbs starts burning, the inductive explanation just fails, whereas the abductive one can supply hints for understanding what happened and for suggesting remedies (Console and Saitta, 2000).

Hence, the two kinds of explanations are very different and distinct, induction firstly aims to provide generalizations, abduction explanations of particular observations. In Console and Saitta's terms (Console and Saitta, 2000), abductive reasoning extends the intension of known individuals (because abducible properties are rendered true for these individuals), without having a genuine generalization impact on the observables (it does not increase their extension)[26].

[25] The role of the causal-hypothetical reasoning is central in modern science: Galileo was already perfectly aware of this fact. He insists that interesting conclusions which reach far beyond experience can be derived from few experiments because "[...] the knowledge of a single fact acquired through a discovery of its cause prepares the mind to understand and ascertain other facts without need to recourse to experiment" (Galilei, 1638): in the case of his study of projectiles, once we know that their path is a parabola, we can derive using only pure mathematics, that their maximum range is 45°. Moreover, Newton says, using the ancient notion of "analysis" (cf. chapter 1, section 1, this book): "By this way of analysis we may proceed from compounds to ingredients, and from motions to the forces producing them; and in general, from effects to their causes, and from particular causes to more general ones, [...] and the synthesis consists in assuming the causes discovered, and established as principles, and by them explaining the phenomena proceeding from them, and proving the explanations" (Newton, 1721, p. 380 ff).

[26] Further results on the interaction and/or integration of abduction and induction in AI complex theory development tasks are given in Michalski (1993) - in terms of coexistence; Dimopoulos and Kakas (1996), and Ourston and Mooney (1994) - in terms of cooperation; O'Rorke, P. (1994), Thompson and Mooney (1994), and Kakas and Riguzzi (1997) - on the role of inducing and learning in abductive theories.

Another interpretation of the distinction between abduction and induction is given by Josephson (2000a). Following his point of view all inferences to the best explanation have to be considered as kinds of "smart" reasoning. In Josephson's terms induction and abduction are not distinct processes: the inductive generalization is a type of inference that points to some best explanation, so it can be considered as a kind of abduction (he agrees with the use of the term abduction according to the second meaning we previously illustrated, the one including generating - or selecting - and evaluating hypotheses). "Smart" inductive generalizations (or inductive hypotheses) do not explain "particular" observations but the frequencies with which the observations emerge, like in the well known AI case of "concept learning from examples": "An observed frequency is explained by giving a causal story that explains how the frequency came to be the way it was. This causal story typically includes both the method of drawing the sample, and the population frequency in some reference class" (Josephson, 2000a).

When inductive hypothesis are "smart" or "good" they are so because they are inferences to the best generalization-explanation of the sample frequencies, so they can be considered as a kind of abduction. As a consequence of his ideas on abduction and induction, Josephson concludes by arguing that the computational programs for inductive generalizations have to be constructed abductively.

3. MODEL-BASED CREATIVE ABDUCTION

3.1 Conceptual change and creative reasoning in science

I have analyzed elsewhere (Magnani, 1999a) some limitations of the sentential models of theoretical abduction in accounting for other reasoning tasks; for example they do not capture 1. the role of statistical explanations, where what is explained follows only probabilistically and not deductively from the laws and other tools that do the explaining; 2. the sufficient conditions for explanation; 3. the fact that sometimes the explanations consist of the application of *schemas* that fit a phenomenon into a pattern without realizing a deductive inference; 4. the idea of the existence of high-level kinds of *creative* abductions I cited above; 5. the existence of model-based abductions (for instance visual and diagrammatic, see section 3.2 below); 6. the fact that explanations usually are not complete but only furnish *partial* accounts of the pertinent evidence (Thagard and Shelley, 1997); 7. the fact that one of the most important virtues of a new scientific hypothesis (or of a scientific theory) is its power of explaining *new*, previously unknown facts:

"[...] these facts will be [...] unknown at the time of the abduction, and even more so must the auxiliary data which help to explain them be unknown. Hence these future, so far unknown explananda, cannot be among the premises of an abductive inference" (Hintikka, 1998, p. 507), observations become real and explainable only by means of new hypotheses and theories, once discovered by abduction.

We will see in the following section that it is in terms of *model-based abductions* (and not in terms of sentential abductions) that we have to think for example of the case of a successful synthesis of two earlier theoretical frameworks which might even have seemed incompatible. The old epistemological view sees Einstein's theory as an attempt to "explain" certain anomalies and facts such as the Michelson-Morley experiment: "The most instructive way of looking at Einstein's discovery is to see it as a way of reconciling Maxwell's electromagnetic theory with Newtonian mechanics [...] it would be ridiculous to say that Einstein's theory 'explains' Maxwell's theory any more than it 'explains' Newton's laws of motion" (Hintikka, 1998, p. 510). This kind of abductive movement does not have that immediate explanatory effect illustrated by the sentential models of abduction: the new framework usually does not "explain" the previous ones but provides a very radical new perspective.

If we want to deal with the nomological and most interesting creative aspects of abduction we are first of all compelled to consider the whole field of the growth of scientific knowledge cited above.

We have anticipated that abduction has to be an inference permitting the derivations of *new* hypotheses and beliefs. Some explanations consist of certain facts (initial conditions) and universal generalizations (that is scientific laws) that deductively entail a given fact (observation), as showed by Hempel in his *law covering model* of scientific explanation (Hempel, 1966): in this case the argument starts with the true premises and deduces the explained event. If T is a theory illustrating the background knowledge (a scientific or common sense *theory*) the sentence α explains the fact (observation) β just when $\alpha \cup T \models \beta$, it is difficult to govern the question involving nomological and causal aspects of abduction and explanation in the framework of the belief revision illustrated in the previous section: we would have to deal with a kind of belief revision that permits us to alter a theory with new conditionals.

We may also see belief change (cf. the previous section) from the point of view of *conceptual change*, considering concepts either cognitively, like mental structures analogous to data structures in computers, or, epistemologically, like abstractions or representations that presuppose questions of justification. Belief revision is able to represent cases of conceptual change such as adding a new instance, adding a new weak rule, adding a new strong

rule (see Thagard, 1992), that is, cases of addition and deletion of beliefs, but fails to take into account cases such as adding a new part-relation, adding a new kind-relation, adding a new concept, collapsing part of a kind-hierarchy, reorganizing hierarchies by branch jumping and tree switching, in which there are reorganizations of concepts or redefinitions of the nature of a hierarchy.

Let us consider concepts as composite structures akin to frames of the following sort:

CONCEPT:
A kind of:
Subkinds:
A part of:
Parts:
Synonyms:
Antonyms:
Rules:
Instances:

It is important to emphasize (1) kind and part-whole relations that institute hierarchies, and (2) rules that express factual information more complex than simple slots. To understand the cases of conceptual revolutions we need to illustrate how concepts can fit together into conceptual systems and what is involved in the replacement of such systems. Conceptual systems can be viewed as ordered into kind-hierarchies and linked to each other by rules. Belief revision is able to represent cases of conceptual change such as adding a new instance, adding a new weak rule, adding a new strong rule (Thagard, 1992), that is, cases of addition and deletion of beliefs, but fails to take into account cases such as adding a new part-relation, adding a new kind-relation, adding a new concept, collapsing part of a kind-hierarchy, reorganizing hierarchies by branch jumping and tree switching, in which there are reorganizations of concepts or redefinitions of the nature of a hierarchy.

Adding new part-relations occurs when in the part-hierarchy new parts are discovered: an example is given by the introduction of new molecules, atoms, and subatomic particles. Thomson's discovery that the "indivisible" atom contains electrons was very sensational.

Adding new kind-relations occurs when it is added a new superordinate kind that combines two or more things previously taken to be distinct. In the nineteenth century scientists recognized that electricity and magnetism were the same and constructed the new concept of electromagnetism. Another case is shown by differentiation, that is the making of a new distinction that generates two kinds of things (heat and temperature were considered the same until the Black's intervention).

The last three types of conceptual change can be illustrated by the following examples. The Newtonian abandon of the Aristotelian distinction between natural and unnatural motion exemplifies the collapse of part of the kind-hierarchy. Branch jumping occurred when the Copernican revolution involved the recategorization of the earth as a kind of planet, when previously it had been considered special, but also when Darwin reclassified humans as a kind of animal. Finally, we have to say that Darwin not only reclassified humans as animals, he modified the meaning of the classification itself. This is a case of hierarchical tree redefinition:

> Whereas before Darwin kind was a notion primarily of similarity, his theory made it a historical notion: being of common descent becomes at least as important to being in the same kind as surface similarity. Einstein's theory of relativity changed the nature of part-relations, by substituting ideas of space-time for everyday notions of space and time (Thagard, 1992, p. 36).

These last cases are the most evident changes occurring in many kinds of creative reasoning in science, when adopting a new conceptual system is more complex than mere belief revision. Related to some of these types of scientific conceptual change are different varieties of *model-based abductions*. In these cases the hypotheses "transcend" the vocabulary of the evidence language, as opposed to the cases of simple inductive generalizations: the most interesting case of creative abduction is called by Hendricks and Faye (1999) trans-paradigmatic abduction. This is the case where the fundamental ontological principles given by the background knowledge are violated, and the new discovered hypothesis transcends the immediate empirical agreement between the two paradigms, like for example in the well-known case of the abductive discovery of totally new physical concepts during the transition from classical mechanics to quantum mechanics.

3.2 Model-based abduction

The last cases of creative reasoning in science we have just illustrated demonstrate the radical *conjectural* character of the new concepts and the incommensurability as regarding previous ones, that is the cases in which "revolutionary" changes happen and the most "counterinductive" acts can become visible. The analysis of model-based conceptual change helps us to study the revolutionary changes of science: different varieties of what I call *model-based* abduction (Magnani, 1999c) are related to some of these types of conceptual change.

Peirce stated that all thinking is in signs, and signs can be icons, indices, or symbols. Moreover, all *inference* is a form of sign activity, where the

word sign includes "feeling, image, conception, and other representation" (*CP* 5.283), and, in Kantian words, all synthetic forms of cognition. That is, a considerable part of the thinking activity is *model-based*. Of course model-based reasoning acquires its peculiar creative relevance when embedded in abductive processes.

For Peirce (Anderson, 1986) a Kantian keyword is synthesis, where the intellect constitutes in its forms and in a harmonic way all the material delivered by the senses. Surely Kant did not consider synthesis as a form of *inference* but, notwithstanding the obvious differences[27], I think synthesis can be related to the Peircian concept of inference, and, consequently, of abduction. After all, when describing the ways the intellect follows to unify and constitute phenomena through imagination Kant himself makes use of the term *rule* "Thus we think a triangle as an object, in that we are conscious of the combination of the straight lines according to a rule by which such an intuition can always be represented" (Kant, 1929, A140, B179-180, p. 182), and also of the term *procedure* "This representation of a universal procedure of imagination in providing an image for a concept, I entitle the schema of this concept" (Kant, 1929, A140-B179-180, p. 182). We know that rules and procedures represent the central features of the modern concept of inference (see also section 4 below). Moreover, according to Peirce, the central question of philosophy is "how synthetical reasoning is possible [...]. This is the lock upon the door of philosophy" (*CP* 5.348), and the mind presents a tendency to unify the aspects which are exhibited by phenomena: "the function of conception is to reduce the manifold of sensuous impressions to unity" (*CP* 1.545).

Most of these forms of constitution of phenomena are creative and, moreover, characterized in a model-based way. Let me show some examples of model-based inferences. It is well known the importance Peirce ascribed to diagrammatic thinking, as shown by his discovery of the powerful system of predicate logic based on diagrams or "existential graphs". As we have already stressed, Peirce considers inferential any cognitive activity whatever, not only conscious abstract thought; he also includes perceptual knowledge and subconscious cognitive activity (Davis, 1972). For instance in subconscious mental activities visual representations play an immediate role.

We should remember, as Peirce noted, that abduction plays a role even in relatively simple visual phenomena. *Visual abduction*[28], a special form of non

[27] For example Peirce considers space and time themselves as products of synthesis and not as forms of intuition, and consequently synthesis can not be seen as a process in time even if Kant would admit that "psychologically" mental phenomena are movements of the mind from one thing to another (Davis, 1972).

[28] Additional considerations concerning this kind of model-based abduction are given in chapter 5, where I also deal with temporal reasoning.

verbal abduction, occurs when hypotheses are instantly derived from a stored series of previous similar experiences. It covers a mental procedure that tapers into a non-inferential one, and falls into the category called "perception". Philosophically[29], *perception* is viewed by Peirce as a fast and uncontrolled knowledge-production procedure. Perception, in fact, is a vehicle for the instantaneous retrieval of knowledge that was previously structured in our mind through inferential processes. Peirce says: "Abductive inference shades into perceptual judgment without any sharp line of demarcation between them" (Peirce, 1955c, p. 304). By perception, knowledge constructions are so instantly reorganized that they become habitual and diffuse and do not need any further testing: "[...] a fully accepted, simple, and interesting inference tends to obliterate all recognition of the uninteresting and complex premises from which it was derived" (*CP* 7.37). Many visual stimuli - that can be considered the "premises" of the involved abduction - are ambiguous, yet people are adept at imposing order on them: "We readily form such hypotheses as that an obscurely seen face belongs to a friend of ours, because we can thereby explain what has been observed" (Thagard, 1988, p. 53). This kind of image-based hypothesis formation can be considered as a form of *visual* (or *iconic*) *abduction*. Of course such subconscious visual abductions of everyday cognitive behavior are not of particular importance but we know that in science they may be very significant and lead to interesting new discoveries (Magnani, et al., 1994; Shelley, 1996). If perceptions are abductions they are withdrawable, just like the scientific hypotheses abductively found. They are "hypotheses" about data we can accept (sometimes this happens spontaneously) or carefully evaluate (cf. also chapter 5, section 1.4)

Peirce gives another interesting example of model-based abduction related to sense activity: "A man can distinguish different textures of cloth by feeling: but not immediately, for he requires to move fingers over the cloth, which shows that he is obliged to compare sensations of one instant with those of another" (*CP* 5.221). This surely suggests that abductive movements have also interesting extra-theoretical characters and that there is a role in abductive reasoning for various kinds of manipulations of external objects (cf. the following chapter, "Manipulative Abduction"). One more example is given by the fact that the perception of tone arises from the activity of the mind only after having noted the rapidity of the vibrations of the sound waves, but the possibility of individuating a tone happens only after having heard several of the sound impulses and after having judged their frequency. Consequently the sensation of pitch is made possible by previous experiences and cognitions stored in memory, so that one oscillation of the air would not produce a tone.

[29] In philosophical tradition perception was viewed very often like a kind of inference (Kant, 1929; Bruner, 1957, Fodor, 1983; Roch, 1983, Gregory, 1987, Josephson, 1994b).

To conclude, for Peirce all knowing is *inferring* and inferring is not instantaneous, it happens in a process that needs an activity of comparisons involving many kinds of models in a more or less considerable lapse of time[30]. This is not in contradiction with the fact that for Peirce the inferential and abductive character of creativity is based on the instinct (the mind is "in tune with nature") but does not have anything to do with irrationality and blind guessing. Hanson (1958, pp. 85-92) perfectly recognizes the model-based side of abductive reasoning (cf. this chapter, section 1.1), when he relates (and reduces) it to the activity of "interpretation" ("pattern of discovery") resorting to the well-known example of reversible perspective figures of *Gestalt* psychology. Unfortunately, this kind of analysis inhibits the possibility of gaining further knowledge about model-based reasoning. I think Hanson is inclined to consider the abductive event as instantaneous and not susceptible to further cognitive and epistemological examination.

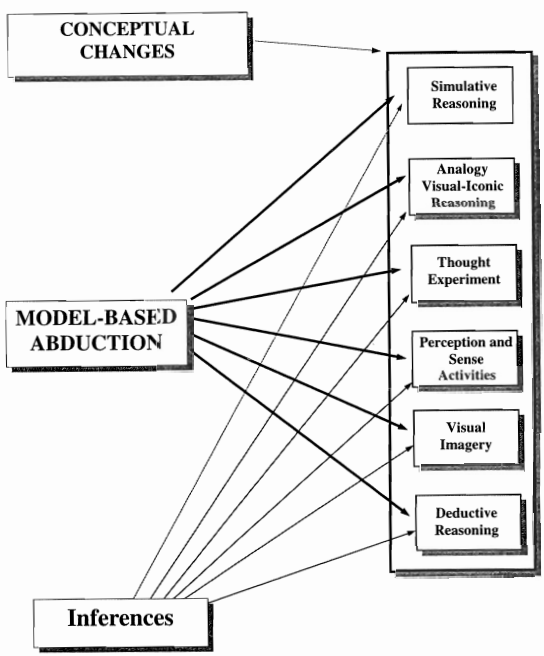

Figure 6. Model-based (theoretical) abduction.

[30] This corresponds to Peirce's "philosophical" point of view, which delineates a very particular meaning of the word "inference", as illustrated above. We have to say that recently very many philosophers and others have accepted that perceptual knowledge is both non inferential and instantaneous (although the latter may be debatable) (Akins, 1996).

All sensations or perceptions participate in the nature of a unifying hypothesis, that is, in abduction, in the case of emotions too:

> Thus the various sounds made by the instruments of the orchestra strike upon the ear, and the result is a peculiar musical emotion, quite distinct from the sounds themselves. This emotion is essentially the same thing as a hypothetic inference, and every hypothetic inference involved the formation of such an emotion (*CP* 2.643);

Also this example surely suggests that abductive movements have interesting extratheoretical effects (see the following chapter)[31]. Human beings and animals have evolved in such a way that now they are able to recognize habitual and recurrent events and to "emotionally" deal with them, like in cases of fear, that appears to be a quick explanation that some events are dangerous. During the evolution such abductive types of recognition and explanation settled in their nervous systems: we can abduce "fear" as a reaction to a possible external danger, but also when affronting a different types of evidence, like in the case of "reading a thriller" (Oatley, 1996).

In all these examples Peirce is referring to a kind of hypothetical activity that is inferential but not verbal, where "models" of feeling, seeing, hearing, etc., are very efficacious when used to build both habitual abductions of everyday reasoning and creative abductions of intellectual and scientific life (see Figure 6).

Following Nersessian (1995a and b), I use the term "model-based reasoning" to indicate the construction and manipulation of various kinds of representations, not necessarily sentential and/or formal[32]. She proposes the so-called cognitive history and philosophy of science approach, which affords a reframing of the problem of conceptual formation and change in science that not only provides philosophical insights but also pays attention to

[31] Considering emotions as abductions, Oatley and Johnson-Laird (1987) have proposed a cognitive theory of emotions largely based on Peircian intuitions. A different aim is pursued by O'Rorke and Ortony (1992, 1994): using a computational tool implemented in PROLOG, AbMaL, and the situation calculus framework, they provide an abductive theory showing how it is possible to construct explanations of emotional states.

[32] See also the recent analysis of the role of models in science given by Giere (1988, 1999), Harris (1999), Raisis (1999), Suarez (1999). For an account on the role of models in the history of recent philosophy of science cf. Bailer-Jones (1999). Zytkow (1999) and Winsberg (1999) describe some aspects of model construction in automated computational systems aimed at reproducing scientific reasoning. On the mediating role of scientific models between theories and the real world cf. Morgan and Morrison (1999). Further, differences in novice and expert reasoning skills in solving scientific problems (cf., e.g., Chi, et al., 1981) provide evidence that skills in modeling is something that develops with learning (Ippolito and Tweney, 1995). Moreover, Nersessian relates model-based reasoning to some aspects of reasoning in terms of "mental models" described by Johnson-Laird (1988, 1993).

the practices employed by real human agents in constructing, communicating and replacing representation of a domain. Common examples of model-based reasoning are constructing and manipulating visual representations, thought experiment, analogical reasoning, but also the so-called "tunnel effect" (Cornuéjols, et al., 2000), occurring when models are built at the intersection of some operational interpretation domain – with its interpretation capabilities – and a new ill-known domain.

We have to remember that visual and analogical reasoning are productive in scientific concept formation too, where the role they play in model-based abductive reasoning is very evident; scientific concepts do not pop out of heads, but are elaborated in a problem-solving process that involves the application of various procedures: this process is a *reasoned process*. Visual abduction, but also many kinds of abductions involving analogies, diagrams, thought experimenting, visual imagery, etc. in scientific discovery processes, can be just called *model-based*. Additional considerations about the intersections between abduction and model-based reasoning (especially in experiment and thought experiment) are illustrated by Gooding (1996): the ability to integrate information from various sources is crucial to scientific inference and typical of all kinds of model-based reasoning also when models and representations are "external", like verbal accounts, drawings, various artifacts, narratives, etc. (on external models and their cognitive role, cf. chapter 3, this book).

We know that scientific concept formation has been ignored because of the accepted view that no "logic of discovery" – either deductive, inductive, or abductive algorithms for generating scientific knowledge – is possible[33]. The methods of discovery involve use of *heuristic* procedures (Peirce was talking of creative abduction as the capacity and the "method" of making good conjectures); cognitive psychology, artificial intelligence, and computational philosophy have established that heuristic procedures are reasoned (see the following section). Analogical reasoning is one such problem-solving procedure, and some reasoning from imagery is a form of analogical reasoning: Holyoak and Thagard (1995) elaborated an analysis of analogical reasoning that encompasses psychological, computational, and epistemological aspects. We have to remember that, among the various kinds of model-based reasoning, analogy received particular attention from the point

[33] It is well-known that Popper (and most of the philosophy of science tradition) confined scientific discovery to the realm of irrationality: "[...] there is not such thing as a logical method of having new ideas, or a logical reconstruction of this process. My view may be expressed by saying that every discovery contains 'an irrational element', or a 'creative intuition', in Bergson's sense" (Popper, 1959, p. 32). This is also the case of the celebrated distinction between "context of discovery" and "context of justification" (Reichenbach, 1938). Rational analysis is only possible within the context of justification (verification, corroboration, falsification).

of view of computational models designed to simulate aspects of human analogical thinking: for example, Thagard, et al. have developed ARCS (Analog Retrieval by Constraint Satisfaction; 1990) and ACME (Analogical Mapping by Constraint Satisfaction; Holyoak and Thagard, 1989), computational programs that are built on the basis of a multiconstraint theory[34].

Hence the role of visual abduction and visual imagery in scientific discovery is very interesting. In chapter 5 (section 1) I will consider visual abduction in everyday reasoning, illustrating a cognitive architecture dealing with image-based explanation.

Finally, by recognizing the role of model-based abduction the analysis of conceptual change can overcome the negative issues that come from the reductionist theory of meaning and from the related incommensurability thesis, and illustrate the various grades of *commensurability* that can be found when dealing with the roles of model-based abduction in science. Nersessian (1998) exploits the representational and constructive virtues of model-based reasoning and makes use of Giere's general idea that "modeling is not at all ancillary to doing science, but central to constructing accounts of the natural world" (1999): she illustrates how model-based abduction can explain that concept transformation and creation involves the construction of fluid and evolving frameworks that guarantee commensurability at many levels.

4. MODEL-BASED HEURISTIC AND DEDUCTIVE REASONING

In chapter 1 I have illustrated how in ordinary geometrical proofs auxiliary constructions are present in term of "conveniently chosen" figures and diagrams where strategic moves are intertwined with deduction. The system of reasoning exhibits a dual character: deductive and "hypothetical". The story of *Meno* dialogue illustrated the role of these strategical moves and their importance in hypothesis generation. These strategical moves correspond to particular forms of abductive reasoning.

As already said in sections 1.2 and 2 the kind of reasoned inference that is involved in creative abduction goes beyond the mere relationship that there is between premises and conclusions in valid deductions, where the

[34] On analogy cf. also the contributions by Kolodner, 1993 (analogy as a form of case-based reasoning in AI), Davies and Goel, 2000 (visual analogy in AI); Gentner, 1982, 1983, 1997 (analogies and metaphors in cognitive science and history of science), Shelley, 1999 (analogy in archaeology). Many theoretical and computational accounts of analogical reasoning have stressed the transfer of relational knowledge. Causal and functional relationships have been the focus of many theories (Holyoak and Thagard, 1995, 1997; Bhatta and Goel, 1997, Falkenhainer, Forbus, and Gentner, 1990; Winston, 1980).

truth of the premises guarantees the truth of the conclusions, but also beyond the relationship that there is in probabilistic reasoning, which renders the conclusion just more or less probable.

On the contrary, we have to see creative abduction as formed by the application of *heuristic procedures* that involve all kinds of good and bad inferential actions, and not only the mechanical application of rules. It is only by means of these heuristic procedures that the acquisition of *new* truths is guaranteed. Also Peirce's mature view on creative abduction as a kind of inference seems to stress the strategic component of reasoning.

Many researchers in the field of philosophy, logic, and cognitive science have sustained that deductive reasoning also consists in the employment of logical rules in a heuristic manner, even maintaining the truth preserving character: the application of the rules is organized in a way that is able to recommend a particular course of actions instead of another one. Moreover, very often the heuristic procedures of deductive reasoning are performed by means of a model-based abduction. A most common example of creative abduction is the usual experience people have of solving problems in geometry in a *model-based* way trying to devise proofs using diagrams and illustrations: of course the attribute of creativity we give the abduction in this case does not mean that it has never been made before by anyone or that it is original in the history of some knowledge.

Hence we have to say that theoretical model-based abductions also operate in deductive reasoning (see Figure 6 above). Following Hintikka and Remes's analysis (1974) proofs of general implication in first order logic need the use of instantiation rules by which "new" individuals are introduced, so they are "ampliative". In ordinary geometrical proofs auxiliary constructions are present in term of "conveniently chosen" figures and diagrams. In Beth's method of semantic tableaux the strategic "ability" to construct impossible configurations is undeniable (Hintikka, 1998; Niiniluoto, 1999)[35].

This means that also in many forms of deductive reasoning there are not trivial and mechanical methods of making inferences but we have to use *models* and *heuristic procedures* that refer to a whole set of strategic principles. All the more reason that Bringsjord (2000) stresses his attention on the role played by a kind of "model based deduction" that is "part and parcel" of

[35] Also Aliseda (1997) provides interesting use of the semantic tableaux as a constructive representation of theories, where abductive expansions and revisions, derived from the belief revision framework, operate over them. The tableaux are so viewed as a kind of reasoning (non-deductive) where the effect of "deduction" is performed by means of abductive strategies.

our establishing Gödel's first incompleteness theorem, showing the model-based character of this great abductive achievement of formal thought[36].

5. AUTOMATIC ABDUCTIVE SCIENTISTS

Paul Thagard (1988) illustrates four kinds of abduction that have been implemented in PI, a system devoted to explaining in computational terms the main problems of the traditional philosophy of science, such as scientific discovery, explanation, evaluation, etc. He distinguishes between simple, existential, rule-forming, and analogical abduction. Simple abduction generates hypotheses about individual objects. Existential abduction postulates the existence of previously unknown objects, such as new planets. Rule-forming abduction generates rules that explain laws. Analogical abduction uses past cases of hypothesis formation to construct hypotheses similar to existing ones. If the pure philosophical task is to state correct rules of reasoning in an abstract and objective way, the use of computer modeling may be a rare tool to investigate abduction in science because of its rational correctness. The increase in knowledge provided by this intellectual interaction is manifest.

Early works on *machine scientific discovery*, such as the well-known Logic Theorist (Newell, Shaw, and Simon, 1957), DENDRAL, in chemistry (Lindsay, et. al., 1980), and AM, in mathematics, (Lenat, 1982), have shown that *heuristic search* in combinatorial spaces is an advantageous and general framework for automating scientific discovery[37]. In these programs abduction is mainly rendered in a sentential way, using rules and heuristics.

There are many ways for identifying a commonality in computational scientific discovery programs that will take a next step beyond the acknowledged general - but weak - framework of heuristic search (cf. also Tweney, 1990). For example, Valdés-Pérez (1999), characterizes discovery in science as the generation of novel, interesting, plausible, and intelligible knowledge

[36] Many interesting relationships between model-based reasoning in creative reasoning and its possible deductive models are analyzed in Meheus (1999), also related to the formal treatment of inconsistencies (cf. chapter 6, section 2.4).

[37] The Journal *Artificial Intelligence* recently devoted a special issue to "Machine Discovery" (91, 1997, - Simon, Valdés-Pérez, and Sleeman, 1997), and so the AAAI Society, that organized a Spring Symposium in 1995 on "Systematic Methods of Scientific Discovery". Classical books where the reader can find the description of the most interesting research and the description of machine discovery programs are Langley, et al., 1987, and Shrager and Langley, 1990. Cf. also Zytkow, 1992 (Proceedings of MD-92 Workshop on "Machine Discovery"), and Colton, 1999 (Proceedings of AISB'99). In 1990 AAAI Society also organized a Spring Symposium on the problem of "automated abduction", devoted to the illustration of many computational programs able to perform various abductive tasks.

about the objects of study. Looking for a common general pattern he analyzes four current machine discovery programs that match those requirements in different ways:

1. MECHEM, which hypothesizes reaction mechanisms in chemistry based on the available experimental evidence (Zeigarnig, et al., 1997)
2. ARROSMITH, which notices connections between drugs or dietary factors and diseases in medicine (Swanson and Smalheiser, 1997)
3. GRAFFITI, which makes conjectures in graph theory and other similar mathematical fields (Fajtlowicz, 1988)
4. MDP/KINSHIP, which delineates the classes within a classification in linguistics (Pericliev and Valdés-Pérez, 1998).

In turn Boden (1991) expecially stresses the distinction between classical programs able to *re-produce* historical cases of scientific discovery in physics (BACON systems and GLAUBER, Langley, et al., 1987), and systems able to perform *new* discoveries (DENDRAL and AM, cited above). Other authors (for example, Schunn and Klahr, 1995, who constructed the program ACT-R) emphasize the distinction between computational systems that address the process of hypothesis formation and evaluation (BACON; PHINEAS, Falkenhainer, 1990; AbE, O'Rorke, Morris, and Schulenburg, 1990; ECHO, Thagard, 1989, 1992; TETRAD, Glymour, et al., 1987; MECHEM), those that address the process of experiment (like DEED, Rajamoney, 1993; DIDO, Scott and Markovitch, 1993), and, finally, those that address both the processes (like KEKADA, Kulkarni and Simon, 1990; SDDS, Klahr and Dunbar, 1988, LIVE, Shen, 1993, and others).

All these AI systems explicitly or implicitly perform epistemological tasks. From the point of view of the task of abduction it is interesting to note that some of them model a kind of sentential creative abduction, others are dealing with model-based creative abduction, and there are also the ones related to model the activity of experiment (that relate to the problem of what I call "manipulative abduction", cf. the following chapter).

Sentential creative abduction. In the first case we have to note that systems like BACON, GLAUBER, built in terms of heuristic search, notwithstanding they perform outputs that can be presented as a fruit of the creative abductive task of re-producing well-known past discovery of physics, they actually execute a selective abduction: starting from given data, they just have to "select" among a pre-stored encyclopedia of mathematical equations capable of explaining the data. Consequently they are similar, because of the epistemology of their architecture, to the computational programs devoted to perform diagnostic reasoning in medicine (cf. chapter 4, this book).

Model-based creative abduction. In the second case the programs are capable of performing model-based abductions: for example by providing causal and analogical reasoning, like the previously cited AbE (theory revision in science), CHARADE (discovery of the causes of scurvy) Corruble and Ganascia, 1997), CDP (discovery of urea cycle, Graßhoff and May, 1995), GALATEA (explanation tasks) (Davies, and Goel, 2000), PHINEAS, that exploits the representational resources of qualitative physics (Forbus, 1984, 1986) to perform analogical reasoning in liquid flow[38]. AbE and PHINEAS explicitly and directly refer to abductive tasks, other programs employ the word induction, even if they are achieving a more complicated task than mere generalization from data[39]. A system that explicitly addresses model-based abduction (the so-called generic modeling) in science is TORQUE (Nersessian, Griffith, and Goel, 1997), devoted to perform tasks of visualizations able to account for various cases of discovery in science (Faraday, Maxwell)[40].

Manipulative abduction. In the third case, when dealing with the simulation of experiment, the computational programs join the area of manipulative abduction. An interesting and neglected point of contention about human reasoning is whether or not concrete manipulations of external objects influence the generation of hypotheses, for example in science: in the following chapter I will delineate the first features of what I call *manipulative abduction* showing how we can find methods of constructivity in scientific and everyday reasoning based on external models and "epistemic mediators". Manipulation of external objects in scientific experiments is a kind of this "epistemic mediator, also exploiting the cognitive resources of human body and its performances. The discovery programs that address the process of experiment constitute the first attempt to automatize these abilities, that could further extend the interest of machine discovery in science also to the whole area of robotics.

It is well known that epistemology is not alone in investigating reasoning. Reasoning is also a major subject of investigation in AI and cognitive psychology. Epistemological theories of reasoning, when implemented in a computer, become AI programs. The theories and the programs are, quite

[38] A system that aims to construct causal hypotheses is TETRAD (Glymour, et al., 1987), but it manipulates numeric data - and not model-based types of reasoning - and is deeply entrenched in a probabilistic framework.

[39] On the ambiguities and relationships between abduction and induction cf. sections 1.2 and 2, this chapter.

[40] Other tools that could be proven useful in the area of abduction and machine discovery come from the field of genetic algorithms and evolving neural networks (cf. Pennock, 1999 and 2000), where creative reasoning is studied improving Darwinian mechanisms described by evolutionary theories, and may be from the very recent so-called DNA computers (Boneh, Dunworth, Lipton, Sgall, 1996).

literally, two different ways of expressing the same thing. After all, theories of reasoning are about rules for reasoning and these are rules telling us to do certain things in certain circumstances. Writing a program allows us to state such rules precisely.

Some philosophers might insist that, between epistemology and cognitive psychology, there is little, if any connection. The basis for such claims is that epistemology is normative while psychology is descriptive. That is, psychology is concerned with how scientists do reason, whereas epistemology with how scientists ought to reason. One of the central dogmas of philosophy is that you cannot derive an ought from an is.

Nevertheless, this kind of ought might be called a "procedural ought". The apparent normativity of epistemology is just a reflection of the fact that epistemology is concerned with rules for how to do something. It would be considerably unreasonable to design a computational model of scientific discovery and reasoning without taking into account how scientists actually reason, what scientists know, and what data scientists can acquire. Nevertheless, the general goal is not the complete simulation of scientists themselves, but rather the achievement of discoveries about the world, using methods that extend human cognitive capacities. The goal is to build prosthetic scientists: just as telescopes are designed to extend the sensory capacity of humans, computational models of scientific discovery and reasoning are designed to extend our cognitive capacities. This cooperation should prove very fruitful from an educational perspective too: reciprocally clarifying both philosophical and AI theories of reasoning will provide new and very interesting didactic tools[41].

[41] Other interesting epistemological and cognitive considerations on the main computational discovery systems are given in Gorman (1998).

Chapter 3

Manipulative Abduction

1. MANIPULATIVE ABDUCTION IN SCIENTIFIC DISCOVERY

The problem of the incommensurability of meaning has distracted the epistemologists from the procedural, extra-sentential and extra-theoretical aspects of scientific practice. Since Kuhn, the problem of translating between languages and of conceptual creativity has dominated the theory of meaning. *Manipulative* abduction (Figure 1) happens when we are thinking *through* doing and not only, in a pragmatic sense, about doing. So the idea of manipulative abduction goes beyond the well-known role of experiments as capable of forming new scientific laws by means of the results (the nature's answers to the investigator's question) they present, or of merely playing a predictive role (in confirmation and in falsification). Manipulative abduction refers to an extra-theoretical behavior that aims at creating communicable accounts of new experiences to integrate them into previously existing systems of experimental and linguistic (theoretical) practices.

The existence of this kind of extra-theoretical cognitive behavior is also testified by the many everyday situations in which humans are perfectly able to perform very efficacious (and habitual) tasks without the immediate possibility of realizing their conceptual explanation. In some cases the conceptual account for doing these things was at one point present in the memory, but now has deteriorated, and it is necessary to reproduce it, in other cases the account has to be constructed for the first time, like in creative settings of manipulative abduction in science. Hutchins (1995) illustrates the case of a navigation instructor that for 3 years performed an automatized task involv-

ing a complicated set of plotting manipulations and procedures. The insight concerning the conceptual relationships between relative and geographic motion came to him suddenly "as lay in his bunk one night". This example explains that many forms of learning can be represented as the result of the capability of giving conceptual and theoretical details to already automatized manipulative executions. The instructor does not discover anything new from the point of view of the objective knowledge about the involved skill, however, we can say that his conceptual awareness is new from the local perspective of his individuality.

In this kind of *action-based* abduction the suggested hypotheses are inherently ambiguous until articulated into configurations of real or imaginated entities (images, models or concrete apparatus and instruments). In these cases only by experimenting we can discriminate between possibilities: they are articulated behaviorally and concretely by manipulations and then, increasingly, by words and pictures.

Gooding (1990) refers to this kind of concrete manipulative reasoning when he illustrates the role in science of the so-called "construals" that embody tacit inferences in procedures involving visual and tactile performances that are often apparatus and machine based. They belong to the pre-verbal context of ostensive operations, that are practical, situational, and often made with help of words, visualizations, or concrete artifacts. The embodied expertise deals of course with an expert manipulation of objects in a highly constrained experimental environment, and is directed by abductive movements that imply the strategic application of old and new (non-conceptual) *templates* of behavior mainly connected with extra-theoretical components.

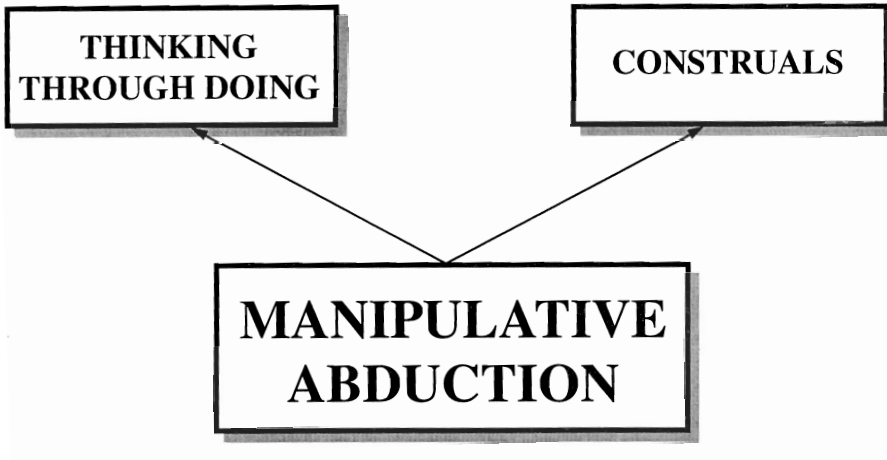

Figure 1. Manipulative abduction.

The hypothetical character of construals is clear: they can be developed to examine further chances, or discarded, they are provisional creative organization of experience and some of them become in their turn hypothetical *interpretations* of experience, that is more theory-oriented, their reference is gradually stabilized in terms of established observational practices. Step by step the new interpretation - that at the beginning is completely "practice-laden" - relates to more "theoretical" modes of understanding (narrative, visual, diagrammatic, symbolic, conceptual, simulative), closer to the constructive effects of theoretical abduction. When the reference is stabilized the effects of incommensurability with other stabilized observations can become evident. But it is just the construal of certain phenomena that can be shared by the sustainers of rival theories. Gooding (1990) shows how Davy and Faraday could see the same attractive and repulsive actions at work in the phenomena they respectively produced; their discourse and practice as to the role of their construals of phenomena clearly demonstrate they did not inhabit different, incommensurable worlds in some cases[1].

Gooding introduces the so called *experimental maps*[2] that are the epistemological two-dimensional tools that we can adopt to illustrate the conjecturing (abductive) role of actions from which scientists "talk and think" about the world. They are particularly useful to stress the attention to the interaction of hand, eye, and mind inside the actual four-dimensional scientific cognitive process. The various procedures for manipulating objects, instruments and experiences will be in their turn reinterpreted in terms of procedures for manipulating concepts, models, propositions, and formalisms. Scientists' activity in a material environment first of all enables a rich perceptual experience that has to be reported mainly as a visual experience by means of the constructive and hypothesizing role of the experimental narratives. Moreover, the experience is constructed, reconstructed, and distributed across a social network[3] of negotiations among the different scientists by means of construals.

These construals aim to arrive to a shared understanding overcoming all conceptual conflicts. As I said above they constitute a provisional creative organization of experience: when they become in their turn hypothetical in-

[1] The theory of manipulative abduction can support Thagard's statement that oxygen and phlogiston proponents could recognize experiments done by each others [8]: the assertion is exhibited as an indispensable requisite for his coherence-based epistemological and computational theory of comparability at the level of intertheoretic relations and for the whole problem of the creative abductive reasoning to the best explanation cited in the previous chapter.

[2] Circles denote concepts (mentally represented) that can be communicated, squares denote things in the material world (bits of apparatus, observable phenomena) that can be manipulated - lines denote actions.

[3] Cf. Minski, 1985 and Thagard, 1997a.

terpretations of experience, that is more theory-oriented, their reference is gradually stabilized in terms of established and shared observational practices that also exhibit a cumulative character. It is in this way that scientists are able to communicate the new and unexpected information acquired by experiment and action.

To illustrate this process - from manipulations, to narratives, to possible theoretical models (visual, diagrammatic, symbolic, mathematical) - we need to consider some observational techniques and representations made by Faraday, Davy, and Biot concerning Oersted's experiment about electromagnetism. They were able to create consensus because of their conjectural representations that enabled them to resolve phenomena into stable perceptual experiences. Some of these narratives are very interesting. For example, Faraday observes: "it is easy to see how any individual part of the wire may be made attractive or repulsive of either pole of the magnetic needle by mere change of position [...]. I have been more earnest in my endeavors to explain this simple but important point of position, because I have met with a great number of persons who have found it difficult to comprehend". Davy comments: "It was perfectly evident from these experiments, that as many polar arrangements may be formed as chord can be drawn in circles surroundings the wire". Expressions like "easy to see" or "it was perfectly evident" are textual indicators inside the experimental narratives of the stability of the forthcoming interpretations. Biot, in his turn, provides a three-dimensional representation of the effect by giving a verbal account that enables us to visualize the setup: "suppose that a conjunctive wire is extended horizontally from north to south, in the very direction of the magnetic direction in which the needle reposed, and let the north extremity be attached to the copper pole of the trough, the other being fixed to the zinc pole [...]" and then describes what will happen by illustrating a sequence of step in a geometrical way:

> Imagine also that the person who makes the experiment looks northward, and consequently towards the copper or negative pole. In this position of things, when the wire is paced above the needles, the north pole of the magnet moves towards the west; when the wire is placed underneath, the north pole moves towards the east; and if we carry the wire to the right or the left, the needle has no longer any lateral deviation, but is loses its horizontality. If the wire be placed to the right hand, the north pole rises; to the left, its north pole dips [...][4].

It is clear that the possibility of "seeing" interesting things through the experiment depends from the manipulative ability to get the correct information and to create the possibility of a new interpretation (for example a

[4] The quotations are from Faraday, 1821-1822, p. 199, Davy, 1821, p. 16, and Biot, 1821, p. 282-283, cited by Gooding, 1990, pp. 35-37.

simple mathematical form) of electromagnetic natural phenomena, so joining the theoretical side of abduction. Step by step, we proceed until Faraday's account in terms of magnetic lines and curves.

It is difficult to establish a list of invariant behaviors that are able to illustrate manipulative abduction in science. As illustrated above, certainly the expert manipulation of objects in a highly constrained experimental environment implies the application of old and new *templates* of behavior that exhibit some regularities. The activity of building construals is highly conjectural and not immediately explanatory: these templates are hypotheses of behavior (creative or already cognitively present in the scientist's mind-body system, and sometimes already applied) that abductively enable a kind of epistemic "doing". Hence, some templates of action and manipulation can be *selected* in the set of the ones available and pre-stored, others have to be *created* for the first time to perform the most interesting creative cognitive accomplishments of manipulative abduction.

Moreover, I think that a better understanding of manipulative abduction at the level of scientific experiment could improve our knowledge of induction, and its distinction from abduction: manipulative abduction could be considered as a kind of basis for further meaningful inductive generalizations. Different generated construals can give rise to different inductive generalizations.

Some common features of these tacit templates (Figure 2) that enable us to manipulate things and experiments in science are related to: 1. sensibility to the aspects of the phenomenon which can be regarded as *curious* or *anomalous*; manipulations have to be able to introduce potential inconsistencies in the received knowledge (Oersted's report of his well-known experiment about electromagnetism is devoted to describe some anomalous aspects that did not depend on any particular theory of the nature of electricity and magnetism; Ampère's construal of experiment on electromagnetism - exploiting an artifactual apparatus to produce a static equilibrium of a suspended helix that clearly shows the role of the "unexpected"); 2. preliminary sensibility to the *dynamical* character of the phenomenon, and not to entities and their properties, common aim of manipulations is to practically reorder the dynamic sequence of events in a static spatial one that should promote a subsequent bird's-eye view (narrative or visual-diagrammatic); 3. referral to experimental manipulations that exploit *artificial apparatus* to free new possibly stable and repeatable sources of information about hidden knowledge and constraints (Davy well-known set-up in terms of an artifactual tower of needles showed that magnetization was related to orientation and does not require physical contact). Of course this information is not artificially made by us: the fact that phenomena are made and manipulated does not render them to be idealistically and subjectively determined; 4. various contingent

ways of epistemic acting: *looking* from different perspectives, *checking* the different information available, *comparing* subsequent events, *choosing*, *discarding*, *imaging* further manipulations, *re-ordering* and *changing relationships* in the world by implicitly *evaluating* the usefulness of a new order (for instance, to help memory)[5].

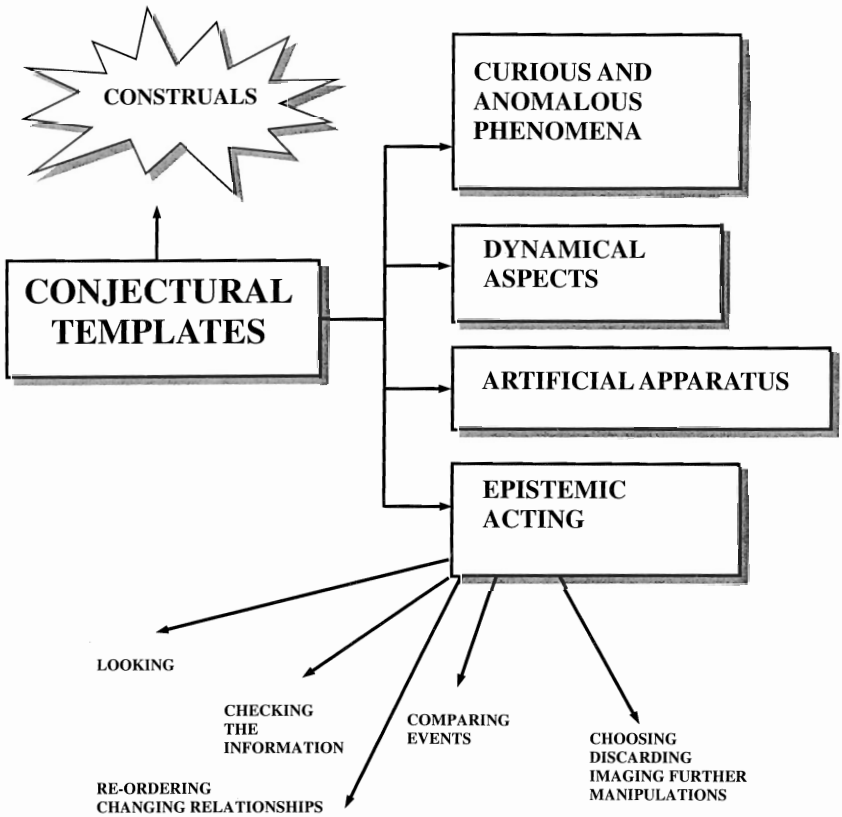

Figure 2. Conjectural templates.

[5] Further aspects of experiment design and its relationship with the problem of communication in science during the transition from the personal to the public domain are given in Gooding and Addis (1999): only a small subset of many observations and measurements performed by individuals of research teams acquire the status of real and public phenomena. Moreover, additional properties of the agent in a scientific experimental setting are described: 1. ability to discriminate between observed results, 2. ability to make judgments about the likelihood of the occurrence of a result, 3. flexibility of the agent's change in perception of the world and his consequent capacity to respond to new information, 4. degrees of competence to build an experiment and observe the results, from novices to experts.

The whole activity of manipulation is devoted to build various external *epistemic mediators*[6] that function as an enormous new source of information and knowledge. Therefore, manipulative abduction represents a kind of redistribution of the epistemic and cognitive effort to manage objects and information that cannot be immediately represented or found internally (for example exploiting the resources of visual imagery)[7].

If we see scientific discovery like a kind of opportunistic ability[8] of integrating information from many kinds of simultaneous constraints to produce explanatory hypotheses that account for them all, then manipulative abduction will play the role of eliciting possible hidden constraints by building external suitable experimental structures. So external well-built structures (Biot's construals for example) and their contents in terms of new information and knowledge, will be projected onto internal structures (for instance models, or symbolic - mathematical - frameworks) so joining the constructive effect of theoretical abduction. The interplay between manipulative and theoretical abduction consists of a superimposition of internal and external, where the elements of the external structures gain new meanings and relationships to one another, thanks to the constructive explanatory theoretical activity (for instance Faraday's new meanings in terms of curves and lines of force). This interplay expresses the fact that both internal and external processes are part of the same epistemic ecology[9].

2. EPISTEMIC MEDIATORS AND MANIPULATIVE REASONING

Recent research, taking an explicit ecological approach to the analysis and design of human-machine systems (Kirlik, 1998), has shown how expert

[6] I derive this expression from the cognitive anthropologist Hutchins (1995), that coins the expression "mediating structure" to refer to various external tools that can be built to cognitively help the activity of navigating in modern but also in "primitive" settings. Any written procedure is a simple example of a cognitive "mediating structure" with possible cognitive aims: "Language, cultural knowledge, mental models, arithmetic procedures, and rules of logic are all mediating structures too. So are traffic lights, supermarkets layouts, and the contexts we arrange for one another's behavior. Mediating structures can be embodied in artifacts, in ideas, in systems of social interactions [...]" (pp. 290-291).

[7] It is difficult to preserve precise spatial relationships using mental imagery, especially when one set of them has to be moved relative to another.

[8] On the role of opportunistic reasoning in design cf. Simina and Kolodner (1995).

[9] It is Hutchins (1995, p. 114) that uses the expression "cognitive ecology" when explaining the role of internal and external cognitive navigation tools. More suggestions on manipulative abduction can be derived by the contributions collected in the recent Morgan and Morrison (1999), dealing with the mediating role of scientific models between theory and the "real world".

performers use action in everyday life to create an *external* model of task dynamics that can be used in lieu of an internal model: "Knowing of not, a child shaking a birthday present to guess its contents is dithering, a common human response to perceptually impoverished conditions". Not only a way for moving the world to desirable states, action performs an epistemic and not merely performatory role: people structure their worlds to simplify cognitive tasks but also in presence of incomplete information or of a diminished capacity to act upon the world when they surely have less opportunities to know. *Epistemic action* can also be described as resulting from the exploitation of latent constraint in the human-environment system. This additional constraint grants additional information: in the example of the child shaking a birthday present he "takes actions that will cause variables relating to the contents of the package to covary with perceptible acoustic and kinesthetic variables. Prior to shaking, there is no active constraint between these hidden and overt variables causing them of carry information about each other". Similarly, "one must put a rental car 'through its paces' because the constraints active during normal, more reserved driving do not produce the perceptual information necessary to specify the dynamics of the automobile when driven under more forceful conditions " (Kirlik, 1998). Moreover, a very interesting experiment is reported concerning short-order cooking at a restaurant grill in Atlanta: the example shows how cooks at various level of expertise use external models in the dynamic structure of the grill surface to get new information otherwise inaccessible.

In this light Powers (1973) studied behavior considering it as a *control of perception* and not only as controlled by perception. Flach and Warren (1995) used the term "active psychophysics" to illustrate that "the link between perception and action [...] must be viewed as a dynamic coupling in which stimulation will be determined as a result of subject actions. It is not simply a two way street, but a circle" (p. 202). Kirsh (1995) describes situations (e.g., grocery bagging, salad preparation) in which people use action to simplify choice, perception, and reduce demands for internal computation through the exploitation of spatial structuring.

We know that theoretical abduction certainly illustrates much of what is important in abductive reasoning, especially the objective of selecting and creating a set of hypotheses (diagnoses, causes, hypotheses) that are able to dispense good (preferred) explanations of data (observations), but fail to account for many cases of explanations occurring in science or in everyday reasoning when the exploitation of the environment is crucial. The concept of manipulative abduction is devoted to capture the role of action in many interesting situations: action provides otherwise unavailable information that enables the agent to solve problems by starting and performing a suitable abductive process of generation or selection of hypotheses.

From the point of view of everyday situations manipulative abductive reasoning exhibits very interesting features (Figure 3): 1. action elaborates a *simplification* of the reasoning task and a redistribution of effort across time (Hutchins, 1995), when we "need to manipulate concrete things in order to understand structures which are otherwise too abstract" (Piaget, 1974), or when we are in presence of *redundant* and unmanageable information; 2. action can be useful in presence of *incomplete* or *inconsistent* information - not only from the "perceptual" point of view - or of a diminished capacity to act upon the world: it is used to get more data to restore coherence and to improve deficient knowledge; 3. action as a *control of sense data* illustrates how we can change the position of our body (and/or of the external objects) and how to exploit various kinds of prostheses (Galileo's telescope - see the following section -, technological instruments and interfaces) to get various new kinds of stimulation: action provides some tactile and visual information (e. g, in surgery), otherwise unavailable; 4. action enables us to build *external artifactual models* of task mechanisms instead of the corresponding internal ones, that are adequate to adapt the environment to agent's needs. Also natural phenomena can play the role of external artifactual models: the stars are not artifacts, but under Micronesian navigator's manipulations of the images of them, the stars acquire a structure that "becomes one of the most important structured representational media of the Micronesian system" (Hutchins, 1995, p. 172). The external artifactual models are endowed with functional properties as components of a memory system crossing the boundary between person and environment (for example they are able to transform the tasks involved in allowing simple manipulations that promote further visual inferences at the level of model-based abduction) (cf. chapter 2, section 3.2 and chapter 5, section 1, this book)[10].

Not all epistemic and cognitive mediators are preserved, saved, and improved, as in the case of the ones created by Galileo at the beginning of modern science (see the following section). For example, in certain non epistemological everyday emergency situations some skillful mediators are elaborated to face possible dangers, but, because of the rarity of this kind of events, they are not saved and stabilized. Hutchins (1995, pp. 317-251) describes the interesting case of the failure of an electrical device, the gyrocompass, crucial for navigation, and the subsequent creation of substitutive contingent cognitive mediators. These cognitive mediators consist of additional computations, redistributions of cognitive roles, and finally, of the dis-

[10] Modeling mechanisms of manipulative abduction is also related to the possibility of improving technological interfaces that provide restricted access to controlled systems, so that humans have to compensate by reasoning with and constructing internal models. New interfaces resources for action, related to task-transforming representations, can contribute to overcome these reasoning obstacles (Kirlik, 1998).

covery of a new shared mediating artifact in terms of divisions of labor - the so-called modular sum that is able to face the situation.

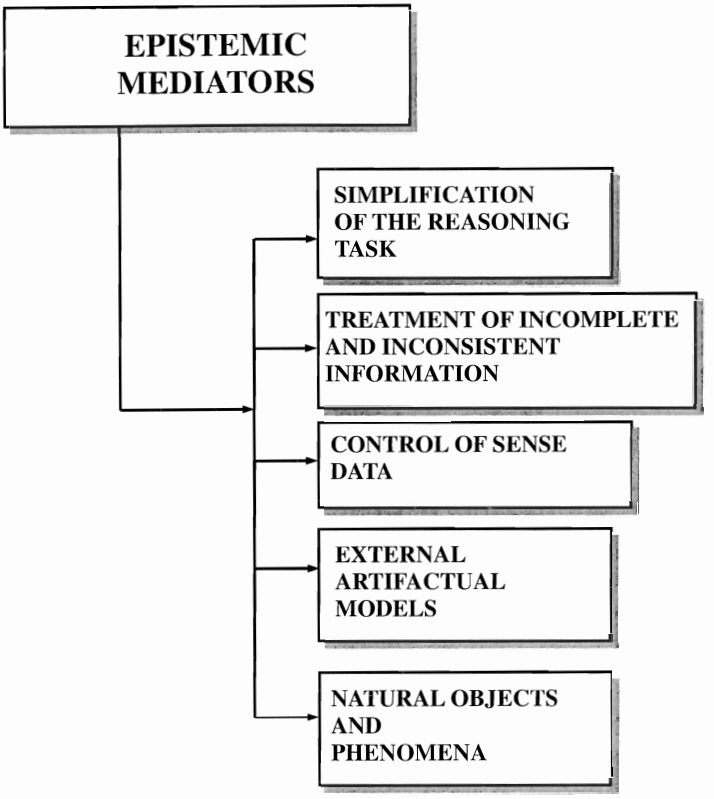

Figure 3. Epistemic mediators.

Finally, we have to observe that many external things that usually are inert from the epistemological point of view can be transformed into epistemic or cognitive mediators. For example we can use our body: we can talk with ourselves, exploiting in this case the self-regulatory character of this action, we can use fingers and hands for counting[11]; we can also use external "tools" like writing, narratives, others persons' information[12], concrete models and

[11] Another example is given by the gestures that are also activated in talking, sometimes sequentially, sometimes in an overlapping fashion.

[12] The results of an empirical research that shows the importance of collaborative discovery in scientific creative abduction and in explanatory activities are given in Okada and Simon, 1997.

diagrams, various kinds of pertinent artifacts. Hence, not all of the cognitive tools are inside the head, sometimes it is useful to use external objects and structures as epistemic devices. We indicated above that Micronesian navigator's stars, that are natural objects, become very complicated epistemic artifacts, when inserted in the various cognitive manipulations (of seeing them) related to navigation.

3. SEGREGATED KNOWLEDGE AND THE "WORLD OF PAPER"

I said that the problem of the incommensurability of meaning has recently distracted the epistemologists from the procedural, extra-sentential and extra-theoretical aspects of scientific practice (section 1 above). This is surprising especially if we consider that the emphasis on concrete manipulative reasoning in case of "construals", that embody tacit inferences in procedures that are often apparatus and machine based, is already clearly granted at the beginnings of modern science.

It is a very common philosophical view to assert that modern science uses experiment to get new information about the world, even if it is not always completely clear the manipulative character of this activity. The new world of the new knowledge has to be totally different from the one merely "of paper" of the Aristotelian tradition. An unbelievable amount of knowledge that was *segregated* had to be released. Accentuating the role of observational manipulations Galileo says:

> The anatomist showed that the great trunk of nerves, leaving the brain and passing through the nape, extended on down the spine and then branched out through the whole body, and that only a single strand as fine as a thread arrived at the heart (Galileo, 1632, p. 63).

Manipulating the cadaver, the anatomist is able to get new, not speculative, information that the Peripatetic philosopher immediately refuses:

> The philosopher, after considering for awhile, answered: "You have made me see this matter so plainly and palpably that if Aristotle's text were not contrary to it, stating clearly that the nerves originate in the heart, I should be forced to admit it to be true" (*ibid.*).

Ipse dixit: no room for the experience. Galileo-Salviati begs of Simplicio: "So put forward the arguments and demonstrations, Simplicio, [...] but not just texts and bare authorities, because our discourses must relate to the sensible world and not the one of paper" (Galileo, 1632, p. 68).

Manipulating observations to get new data, and "actively" building experiments, like the famous one from the leaning tower, sometimes with the help of artifacts, is the essence of the new way of knowing. Galileo says: "All these facts were discovered and observed by me many days ago with the aid of a spyglass which I devised, after first being illuminated by divine grace. Perhaps other things, still more remarkable, will in time be discovered by me or by other observers with the aid of such an instrument" (Galileo, 1610, p. 28). Attaching a scale marked with equally spaced horizontal and vertical lines to his telescope, and manipulating objects "idealizing" them and not considering interesting and non influential factors, Galileo was able to record the daily histories of the four "starlets" accompanying Jupiter and to show that the data was consistent with the abduction that the starlets were indeed moons orbiting Jupiter with a constant period.

With Galileo's achievements, we observe that human scientific thinking is related to the manipulation of a material and experimental environment that is no longer natural. Knowledge is finally seen as something cognitively distributed across scientists, their internal "minds", and external artifacts and instruments. Experiments and instruments embody in their turn external crystallization of knowledge and practice. Modern science is made by this interplay of internal and external. Bacon too was very clear about this distribution of epistemic tasks:

> Those who handled sciences have been either men of experiment or men of dogmas. The men of experiment are like the ant, they only collect and use; the reasoners resemble spiders, who make cobwebs out of their own substance. But the bee takes a middle course: it gathers its material from the flowers of the garden and of the field, but transforms and digests it by a power of its own (Bacon, 1620, p. 52).

An immediate consequence of Galileo's and Bacon's ideas is the critique of the authority, that advocated the knowledge relevance of a "world of paper", mainly internal from the cognitive point of view. Gooding observes: "It is ironical that while many philosophers admire science because it is empirical as well as rational, philosophical practice confines it to the literary view that Galileo rejected" (1990, p. xii). Galileo's "book of nature" and his systematic use of the telescope are the revolutionary *epistemic mediators*[13] that characterize the cognitive power of the new way of producing intelligibility.

Changes in the modalities of distributing epistemic assignments are never without costs. We can just remark, even if well-known, that Galileo's new management of information and knowledge by means of inventing and stabilizing these mediators was not without individual and social costs. Because of the new knowledge provided, *Dialogue* was prohibited, and he was sen-

[13] Together with the exploitation of mathematical models.

tenced (1633 - the admonition is of 1616) to life imprisonment by the Holy Office with the added task of having to recite once a week for three years the seven penitential psalms. He read his abjuration and was released to the custody of the Archibishop of Siena; his daughter, Sister Maria Celeste was given permission to recite the psalms in his stead (van Helden, 1989). The deterioration of the scientific climate and the decline of telescopic astronomy in Italy were the obvious immediate consequences. Notwithstanding the problems, these new epistemic mediators, that are at the roots of the tradition of scientific knowledge, were preserved, saved, and subsequently improved.

It is well known that recent philosophy of science has paid great attention to the so-called theory-ladenness of scientific facts (Hanson, Popper, Lakatos, Kuhn): in this light the formulation of observation statements presupposes significant knowledge, and the search for important observable facts in science is guided by that knowledge. It is absolutely true that theory is able to lead us to abduce new fact, but we cannot forgot that a lot of new information is reached by observations and experiments, as fruit of various kinds of artifactual manipulations. Robert Hooke, using microscope to look at small insects, with practical interventions illuminated his specimens from different directions to establish which features remained invariant under such changes and discovered that some disagreements about data were apparent (Chalmers, 1999, p. 22). Galileo did not have a theory about Jupiter's moons to test when he used his telescope, but the manipulations of the new technology offered a lot of new information. In these cases it is only later that theory is able to contribute new meanings to experimental results.

Following the so-called "new experimentalism" (Ackermann, 1989) experiment has a "life of its own" (Hacking, 1983), independent of theory. Hacking declares:

> Experimental work provides the strongest evidence for scientific realism. This is not because we test hypotheses about entities. It is because entities that in principle cannot be "observed" are regularly manipulated to produce new phenomena and to investigate other aspects of nature. They are tools, instruments not for thinking, but for doing (p. 262).

We are even able to manipulate the "philosopher's favorite theoretical entity", the electron, and it is only in the early stages of our discovery of that entity, that we may merely test the hypothesis that it exists. We already said that a great part of the recent philosophy of science is theory-dominated: data is always considered as theory-laden. All histories of scientific facts are written, in this light, to emphasize theory and forget the experimental and technological aspects of research: experiments do not have an autonomous significance and the explanation of their characteristics, aims and results is made in terms of theoretical issues unknown to the experimenter. For in-

stance: the experiment is considered significant only as a means to test a theory under scrutiny. Hacking provides an interesting analysis of Lakatos' treatment of Michelson's experiment: Hacking's description of this experiment tells us that it does not pursue any programme Lakatos writes about and it has a relative autonomy as regards theory. Classic positivism, pragmatism and kantism, the philosophies of science of Carnap, Popper, Lakatos, Feyerabend, Putnam, van Fraassen and others are characterized by a "single-minded obsession with representation and thinking and theory, at the expense of intervention and action and experiment" (p. 131).

Contrarily to the recent epistemological tradition, we have to follow Hacking and stress the attention on manipulative abduction and epistemic mediators also from the cognitive point of view. Creating an external cognitive support is very important to increase the possibility to get new information, to extend scientific knowledge, but also to improve and simplify many kinds of reasoning. Scientific thinking, like everyday thinking, has not to be viewed only like an internal speculative cognitive process, which occurs in a detached contemplation.

Hacking considers also the problem of realism by analyzing what we can use to intervene in the world to affect something else, or what the world can use to affect us. He shows, with the help of many interesting and sophisticated laboratory examples - some of them full of historical interest - that the significance of experiments sometimes has little to do with theory and representation. Entities whose causal powers are well understood are used as tools to investigate (and to intervene in) nature:

> Understanding some causal properties of electrons, you guess how to build a very ingenious complex device that enables you to line up the electrons the way you want, in order to see what will happen to something else. [...] Electrons are no longer ways of organizing our thoughts or saving the phenomena that have been observed. They are ways of creating phenomena in some other domain of nature. Electrons are tools (p. 263).

Concepts become tools endowed with absolutely unexpected outcomes. The experimentalists use various strategies for establishing the experimental effects without any recourse to theory. These strategies correspond to the expert manipulation of objects in a highly constrained experimental environment, we said directed by abductive movements that imply the application of old and new extra-theoretical *templates* of expert behavior. As possible creative organization of experience some of them become in their turn hypothetical *interpretations* of experience, that is more theory-oriented, their reference is gradually stabilized in terms of established observational practices. Step by step the new interpretation relates to more "theoretical" modes

of understanding (visual, diagrammatic, symbolic, conceptual, simulative), closer to the constructive effects of theoretical abduction.

In this light it is not surprising that Mayo (1996), in her defense of experimentalism, has recently stressed attention to the possibility of delineating progress in science in terms of accumulation of experimental knowledge and expertise. She adds more arguments to the thesis of autonomy of experimental results illustrating many examples where the experiments are shown not as merely related to confirmation and falsification. In some cases they not only serve as a falsification of the assertion, but also to delineate new effects and ideas not previously known; moreover, they can bear on the comparison of radically different theories[14]: to resume, they can trigger revolutionary creative abductions, enabling us to learn from errors. To exemplify the positive role played by errors Mayo illustrates the famous case of the observation of the questionable features of Uranus's orbit that created problems for Newtonian theory: the detection of the source of this difficulty led to the discovery of Neptune.

4. NON-CONCEPTUAL AND SPATIAL ABILITIES

The importance of manipulative abduction can also be understood by considering the recent tradition of cognitive research in *dynamical systems*. The importance and the validity of this tradition is controversial in the cognitive science community, yet I think it can provide many interesting suggestions to further develop the abductive cognitive aspects of external and bodily epistemic mediators. It is well-known that cognitive science has been dominated by the approach that considers cognition as an operation of a special mental computer, situated in the brain[15]. Sensory organs discharge representations of the state of the environment to the mental computer. The system processes a specification of appropriate thought, reasoning or actions. In the last case it is the body which carries out the action. Representations are considered like static structures of symbols, and cognitive procedures are sequential transformations from one structure to the next.

Following the cognitive approach suggested by research in dynamical systems, we immediately learn that cognitive processes and their context have to be considered as unfolding "continuously and simultaneously in real time" (van Gelder and Port, 1995). This approach (Clark, 1997) tries to ex-

[14] I already stressed (section 1, above) the role played by construals of phenomena to overcome the problem of incommensurability indicating that they make able the sustainers of different worlds of knowledge to share something common.

[15] It is the so-called CRUM (Computational-Representational Understanding of Mind) illustrated and criticized by Thagard (1996).

plain how a cognitive system can generate its own change, displaying its self-organizing character. Hence, a cognitive system is not a computer, but a dynamical system, a whole system including the nervous system, body (with its movements, feelings, and emotions), and environment: every cognitive process takes place in real biological hardware and this *embeddedness*, and the various "interactions" involved, become its central aspect.

Many studies in the area are related to the way infants acquire simple body skills and actions, showing how thought is embodied: thought grows from actions and from control of the body, as already guessed and partially described by Piaget (1952), who illustrated the role of sensory-motor period of here-and-now bodily existence. For example we have a felt embodied understanding of bilateral symmetry that can be considered the basic experiential level from which the highest levels of human art and language are developed. Experiments show that exploration of the environment and control of body in children is the fundamental step that favors the cognitive constructions of hidden embodied templates and patterns of kinematic skills but also of speech abilities: "what infants sense and what they feel in their ordinary looking and moving are teaching their brain about their bodies and about their worlds" (Thelen, 1995).

When explained in dynamic terms (Saltzman, 1995), the activity of perceptual motor category formation can be easily seen as highly abductive and foundational for all cognitive development. Of course, the interactions between the body and the environment are not the only form of cognitive mediation with the world. It is obvious that social communication too (and of course manipulations of the conditions of possibility of this communication - cf. the previous section) provides rich information to many of our senses.

Other research that involve the problem of embodiment come from the field of robotics (Chrisley, 1995), where the study of human cognition emphasizes the role of acts in real-time and real-space environments, going beyond the computer/brain model to expand the analogy to the robot/body. Robotic computation can help to learn how embeddedness and embodiment of non-conceptual abilities in an intentional system can be delineated and understood (cf. also Dorigo and Colombetti, 1998).

A very interesting kind of formation of non-conceptual hidden templates, patterns, and skills, is illustrated by the generation of *spatial abilities* in children, adults, non-human mammals and birds. This investigation also challenges some certainties of the propositional view of the mind. Non-conceptual spatial patterns are abductively formed in cognitive situations that are highly integrated with external objects, environment, and body movements (Foreman and Gillett, 1997). Many cognitive cases are studied using traditional experimentation in psychology and ethology, but also taking advantage of a neurobehavioral approach: 1. egocentric ways of encod-

ing space in comparison to the allocentric ones, capable of guaranteeing navigational spatial skills, 2. spatial mapping in people, 3. homing behavior of various mammals and birds (Bovet, 1997), 4. use of landmarks and/or gestures in orientation and perceptual localization (to reach external objects) (Bloch and Morange, 1997), 4. several ways - landmark, route, and survey based - of children's way finding (Blades, 1997), 5. people navigating in large scale environments (Gärling, Selart, and Böök, 1997).

Finally, it is interesting to note that in mammalian species the sensory-motor activity of exploration is displayed in the presence of novel and/or unexpected events and situations. This kind of investigatory behavior seems to be part of a general process of knowledge. Also environment and interactive "punishments" in cases of non-human animals and robots spatial explorations can be cognitively seen as playing the role of contradicting some expected outcomes and as an important step that encourages and supports the emergence of more or less stable spatial templates. In chapter 6, I will illustrate the importance of contradictions, novelty, and unexpected findings in various kinds of theoretical abductions in reasoning.

We can guess that some abduced patterns and templates acquired and formed during exploration constitute intermediate steps between body-centred non-conceptual abilities and further representational and more general and abstract frameworks in humans as well, like for example cognitive maps and diagrams and geometrical knowledge and generalizations[16]:

> It can be reasonably hypothesized that during exploration, feedback arising from the trajectory itself (vestibular, muscular, and other information) is matched with the ever-changing visual scenes of the environment. This matching would inform the subject that it is moving in a stable environment whose properties are invariant whatever their perceptual appearance (Thinus-Blanc, Save, and Poucet, 1997).

[16] I this section I have very quickly illustrated some important points that I consider of interest for delineating some more aspects of manipulative abduction. More details and a broader treatment will be found in a book on philosophical and cognitive problems of geometry I am preparing (Magnani, 2000).

Chapter 4

Diagnostic Reasoning

After having described the first philosophical features of abductive reasoning in chapter 2 I introduced an epistemological model (ST-MODEL: Select and Test Model) of medical diagnostic reasoning which can be described in terms of abduction, deduction and induction (section 1.2).

In this chapter I will exploit this model to describe the different roles played by these basic inference types in developing the various kinds of *medical reasoning* (section 1); I will relate it to some *cognitive models* of medical reasoning (section 2), and, finally, I will provide an abstract representation - an epistemological architecture - of the control knowledge embedded in a *medical knowledge-based system* (KBS) (section 3), and a succinct description of the computational program NEOANEMIA (section 4). In my opinion the controversial status of abduction is related to a confusion between the epistemological and cognitive levels, and to a lack of explanation as to why people sometimes deviate from normative epistemological principles.

Exploiting the epistemological model in order to design the general inferential behavior (control knowledge) of a medical KBS leads to creating a more complex one with an ontological level dealing with the entities and relationships belonging to the underlying domain knowledge. Different ontologies express diagnosis, therapy planning and monitoring but the three tasks can be executed by a single inference process in terms of abduction, deduction, and induction, in order to solve problems (section 1).

The aim of this chapter is also to emphasize the distinction between *basic medical science* (and reasoning) and *clinical science* (and reasoning) in order to illuminate some basic philosophical and cognitive issues in medical education. The Kunhian concept of exemplar refers to the field of growth of scientific knowledge and in this sense is related to the "anti-theoretical" empha-

sis on problem-solving performance. In cognitive science this (and similar) types of postpositivistic objections to the formalistic excess of the neopositivistic tradition are exploited to stress the relevance of the distinction between theories and their domains of application. This objection is exploited to stress the difference between established bodies of scientific knowledge and their processes of discovery and/or application and, in medical knowledge, between clinical reasoning (situated, concerned with attributes of people) and basic science reasoning (unsituated, concerned with attributes of entities such as organs, bacteria, viruses).

Exploiting the theoretical consequences of the analysis of the previous topics I will try to answer some questions: What is the role of problem-solving in teaching and learning, as different from conventional basic science-centred education? Is it relevant, in medical education, an epistemological and logical awareness of the main methodological topics? Finally, the analysis of the significance of abduction in a unified epistemological model of medical reasoning is exploited to individuate the proper ontological level dealing with the entities and relationships belonging to the dynamism of the underlying domain knowledge (for instance biomedical physics) and the consequences for *medical education* (sections 5 and 6).

1. IS MEDICAL REASONING ABDUCTIVE?

As already anticipated in chapter 2 (section 1.2), in accordance with the ST-MODEL, medical reasoning may be broken down into two different phases: first, patient data is abstracted and used *to select hypotheses*, that is hypothetical solutions of the patient's problem (selective abduction phase); second, these hypotheses provide the starting conditions for forecasts of expected consequences which should be compared to the patient's data in order to *evaluate* (corroborate or eliminate) those hypotheses which they come from (deduction-induction cycle).

In the case of medical KBSs[1] the epistemological architecture which exploits the abduction-deduction-induction cycle (ST-MODEL) starts with an abstraction of the data which characterizes the problem to be solved (diagnosis, therapy, monitoring). An abstraction can be considered as a process of structuring incoming data in a smaller set of entities, according to the kind of medical knowledge available and the features of the problem at issue. The efficacy of such operations depends on accumulated expertise, which determines the organization of personal knowledge so that problems can be easily recognized and stated in a way that guarantees their solution by efficient use

[1] Deriving from the AI tradition established for example by Shortliffe, 1976.

of available knowledge. Clancey's well-known distinction between definitional abstraction, qualitative abstraction and generalization and certain other aspects of abstraction are illustrated by Stefanelli and Ramoni (1992), and Ramoni, et al. (1992).

Patel, Evans, and Kaufmann (1989), characterizing physicians' performances in experimental research (in this case on diagnostic expertise), suggest a more pragmatic and active structure organized in a multi-level problem-oriented framework (previously developed in a broader model by Evans and Gadd (1989). They identify appropriate abstractions as "units of knowledge" to code influentially doctor-patient interaction. So clinical knowledge is hierarchically organized from observation to findings to facets (diagnostic components) to diagnosis. Observations are units of information considered as potentially relevant according to the features of the problem-solving context. Findings are composed of sets of observations that are relevant in a diagnostic context. Facets are clusters of findings that are suggestive of diagnostic components. Specific combinations of facets lead to a diagnosis. The aim is to capture "how a clinician identifies problem-specific cues, concludes findings and derives meaning from higher-order relations in the data" (Patel and Groen, 1991).

Selective abduction simply involves guessing a set of hypotheses starting from problem features identified by abstraction. Once hypotheses have been selected, they need to be ranked (Figure 1) so as to plan the evaluation phase by first testing a certain preferred hypothesis. As already suggested by Peirce, there are multiple criteria for ranking; in the medical domain they may be parsimony, danger, cost, curability, and so on, that are chosen according to the specific knowledge context. The worth of a hypothesis to be tested first is, of course, connected to epistemic and pragmatic collections of reasons that trace back to belief in its truth and general relevance for medical action. Of course in the case of diagnosis (where the role of "explanation" is dominant) the epistemic reasons will be dominant, whereas, in the case of therapy, epistemic reasons will be intertwined with pragmatic and ethical reasons, which will play a very important role.

The *deduction-induction phase* deals with the actual process of hypotheses' evaluation. Deduction is connected with prediction. Once a hypothesis about a patient is established (for example, a diagnosis or a type of therapy), certain predictions derived at a time t_1 (the presence of a certain symptom, the development of certain consequences, estimates of a particular evolution) can be revised at t_2: the conclusions are defeasible, that is they may be retracted when new information establishing that the condition "all other things being equal" has been disproved.

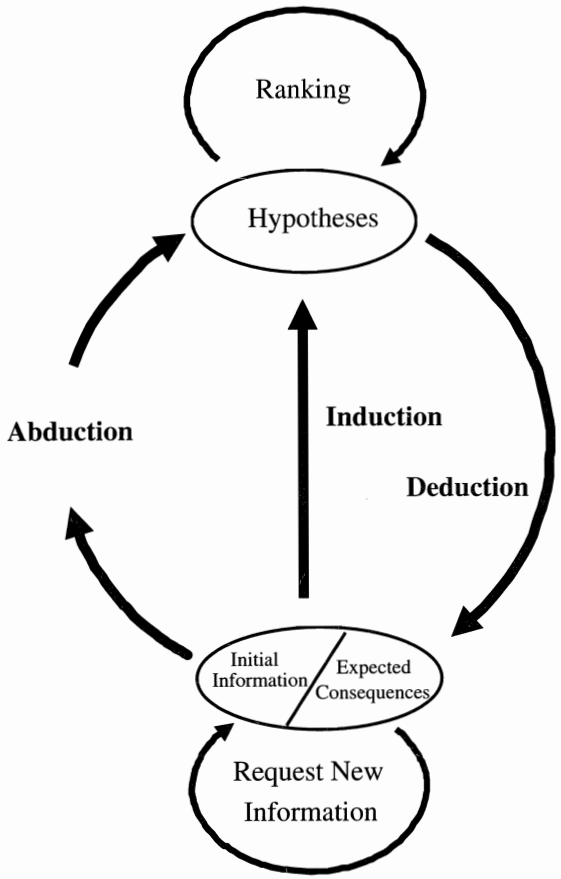

Figure 1. The epistemological model of hypothetical reasoning.

Induction (not used here to mean an ampliative process of the generalization of knowledge) corroborates those hypotheses whose expected consequences turn out to be consistent with the patient's data and refutes those which fail this test. It is important to remember, as I observed above in chapter 2, that, in an ideal situation, it would be necessary to achieve the best explanation by evaluating the uneliminated competing hypotheses so as to test their explanatory power. Induction is the final testing of an abduced hypothesis: by completing the whole cycle of the epistemological model it produces the best explanation.

If new information suggests hypotheses not previously considered, a new cycle starts. The cyclic nature of the epistemological model stresses its non-

monotonic character, as stated above, and this is even more the case for medical reasoning.

Diagnosis and therapy planning (but also patient monitoring) can be executed by an instance of the epistemological model described above, as shown in Figure 3 (chapter 2) and in Figure 2. Of course the ontologies involved are different: there are diagnostic hypotheses, manifestations etc. in diagnostic reasoning; therapies, therapeutic problems and so on in therapy planning; alarms, critical conditions, emergency actions and so on in monitoring.

Diagnosis is the first task to be executed in medical reasoning. It starts from patient data that is abstracted into clinical features to be explained. Then, selective abduction generates plausible diagnostic hypotheses. Starting from the highest ranked hypothesis, deduction shows the findings that are expected if this hypothesis is true. Thus new laboratory or clinical examinations can be requested to verify unobserved expectations. Finally induction establishes whether hypotheses can be confirmed or refuted, or whether they are worth testing further, depending on how closely the observed findings compare with expectations. Induction is the final testing of an abduced diagnostic hypothesis and involves the whole cycle of the epistemological model produces the best diagnostic explanation. Furthermore, this inference type deals with the termination of the diagnostic process: it decides whether a satisfactory explanation of the patient's state has been achieved.

In *therapy planning*, that starts with the observational data and a diagnosis, when available, selective abduction generates plausible (i.e. potentially useful) treatments. Such a task not only involves mapping continuous values of clinical variables into meaningful categorical propositions, but also, and more important, deriving a restricted set of critical aspects of the patient's condition which can be immediately interpreted as a list of crucial targets of the therapy.

The successive abduction takes the list of therapeutic problems and infers a presumptive list of therapies which includes those treatments that deserve consideration as potentially useful in handling those problems. Far from being definitive, elements of this list are considered just as hypotheses, needing further more focused analysis and testing. In fact, reasoning proceeds with a ranking phase, which usually establishes priorities over the current list of treatments.

The testing phase involves the deductive-inductive inference. Deduction here consists in focusing on single treatments of the list, or pairs of them, in order to perform a more thorough evaluation of their appropriateness for the patient at hand. This usually involves making predictions in order to estimate possible consequences of the treatments on the clinical course of the specific patient.

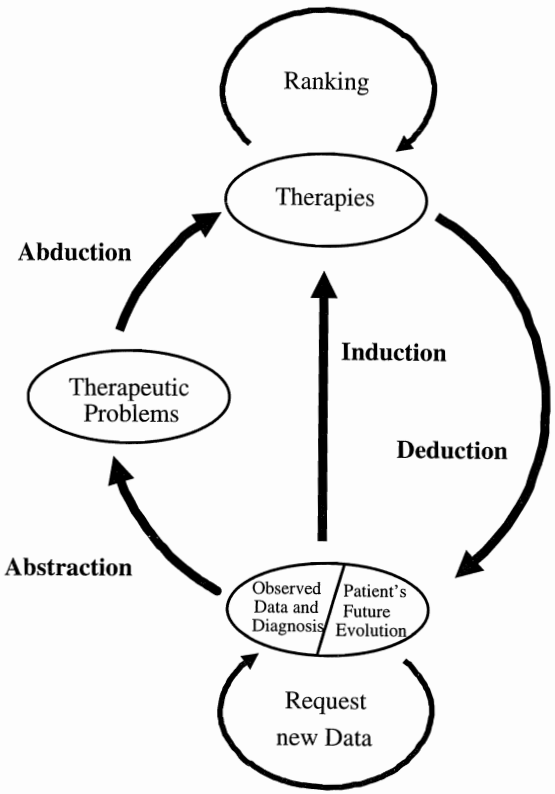

Figure 2. The epistemological model of therapeutical reasoning.

On the basis of these predictions, some therapies may be excluded from the initially abduced set. Both in diagnosis and in therapy planning, if new information suggests hypotheses (diagnostic or therapeutic) not yet considered, a new cycle takes place.

As stressed above, this requires nonmonotonic reasoning. New data usually trigger further cycles, until the list of treatments is reduced to the point of providing a helpful advice.

Monitoring, that is observing and controlling the course of a patient's condition, is the best strategy to test if the planned action proves to be properly effective. It is essential to answer here the following question: does monitoring represent a different generic task from diagnosis and therapy planning? Without taking into consideration very specific situations, diagnosis can be considered as the task of achieving the best explanation of patient's condition, therapy planning the best action to perform in order to im-

prove patient's condition, and monitoring the best strategy to verify if the planned action proves to be really effective. To this aim, a KBS should be able to predict the course of a patient's condition under the combined action of diagnosed disorders and selected therapy. This ability should also characterize the action of an expert.

From an epistemological point of view, monitoring may be described by the ST-MODEL. If the selected therapy works and the patient is responding appropriately, according to the specific patient model used, then therapy is continued or the patient is released from treatment (testing phase). If the therapy did not work or if unusual findings arise, then further assessment is necessary (selecting phase). As a result of monitoring, previous diagnosis and therapy planning tend to be either confirmed or rendered questionable. In the former case monitoring implies continuous cycling between deduction and induction, while in the latter case diagnosis and/or therapy planning may need to be executed again, so requiring abductive inferences starting from the new patient's condition. The way in which monitoring is another instance of the epistemological model (ST-MODEL) is illustrated in Ramoni, et al. (1992) and Stefanelli and Ramoni (1992).

2. COGNITIVE MODELS

AI research has developed many computational tools for describing the representation and processing of information. Cognitive psychologists have found these tools valuable for developing theories about human thinking and for their experimental research. Notwithstanding this, the study of methods of inquiry falls primarily within the province of philosophers of science rather than of scientists themselves, principally because these issues are normative rather than descriptive. To escape relativism, epistemology is usually considered as the normative theory of objective knowledge, and thus does not need to take into account what psychology determines as the nature of individuals' belief systems. Logic and epistemology are concerned with how people ought to reason, whereas psychology is supposed to describe how people do think[2].

Empirical studies of cognitive psychology are descriptive: they are dedicated to the investigation of mental processes and are concerned with normative issues only in order to characterize people's behavior relative to assumed norms. AI, when examined as cognitive modeling, is normally de-

[2] On the relationships and differences between reasoning in "real" human agents and normative logical theories cf. Evans, 1982.

scriptive: only when it is concerned with improving on people's performances does it become involved with what is normative.

Epistemology, AI and cognitive psychology can be used together to develop models that explain how humans think (Thagard, 1988, 1996):

> A psychological model should be more than internally coherent [as is the case with computer simulation]: we want it to account for experimental data about how people think. But sometimes, if a model is complex, it is not easy to see what its consequences are. Cognitive models, like many models in the social sciences, often postulate many interacting processes. The computer program enables a researcher to see whether the model has all and only the consequences that it was expected to have. Comparison of these consequences against experimental observations provides the means of validating the model in much greater detail than pencil-and-paper calculations might allow (Thagard, 1988, p. 6).

I would like to illustrate the relationships and compatibility between my epistemological framework of medical reasoning and certain cognitive models of physicians' reasoning.

As we have seen in chapter 2, section 1.2, if abduction is considered as inference to the best explanation, abduction is epistemologically classified not only as a mechanism for selection (or for discovery), but for justification too. In the latter sense the classical meaning of abduction as inference to the best explanation (for instance in medicine, to the best diagnosis or the best therapy) is described in my epistemological model by the complete cycle abduction-deduction-induction (Josephson, et al., 1986). Nevertheless, as we have seen, abduction can be considered simply as a mechanism for production of plausible hypotheses, and this is the case with my epistemological model. The need for a methodological criterion of justification is caused by the fact that an abduced hypothesis that explains a certain puzzling fact should not be accepted because of the possibility of other explanations. Having a hypothesis that explains a certain number of facts is far from a guarantee of being true.

I think this controversial status of abduction is related to a confusion between the epistemological and cognitive levels, and to a lack of explanation as to why people sometimes deviate from normative epistemological principles. An analysis of the differences between epistemological and cognitive levels would help to clarify the issue.

From an *epistemological* point of view, abduction as inference to the best explanation involves the deduction-induction cycle of testing by means of multi-dimensional criteria of evaluation. Abduction as inference that provides a *possible* explanation of some puzzling phenomenon, is only a mechanism of discovery (or in medical diagnosis, of selection). In this latter

sense abduction is the "wild hunch" that may either be a brilliant breakthrough or a dead-end: nevertheless it implies *uncertainty*, which can be removed or reduced only by testing the implications of selected diagnostic hypotheses against the available data. From an *empirical* point of view, for instance in the case of experimental research on the behavior of physicians, there is an external criterion of truth: the correctness of a diagnostic conclusion is already known (the best diagnosis) in relation to a particular condition, and this is compared to observations of a physician's performance.

There exist many possibilities and many diagnostic performances are found: physicians make correct (best), or wrong diagnoses both by an abduction/deduction-induction cycle of testing (abduction considered as inference to the best explanation according to the complete cycle of my epistemological model), and by selective abduction (without the testing cycle).

The empirical regularities established by Patel and Groen (1991) from research on expert-novice comparisons illustrate, among other things, the role of *forward reasoning* and *backward reasoning* in medical diagnosis. Because of the revealed independence of recall phenomena from diagnostic accuracy (diagnostic accuracy is developmentally monotonic whereas recall is nonmonotonic; the development of expertise is not related to the development of increasingly better representations) the main results of this research lead to a rejection of (1) the theory of medical diagnosis as pattern recognition and (2) the theory of diagnostic expertise based on a set of production rules. Here is a résumé of certain important results from this empirical research into the various kinds of diagnostic reasoning. The first kind of reasoning - forward - as a *strong* problem-solving method that requires a great deal of relevant knowledge, is error-prone in the absence of adequate domain knowledge. The second - backward -, as a *weak* method, is used when domain knowledge is inadequate or when relevant prior knowledge is lacking. In my epistemological model, forward reasoning (knowledge-based heuristic search - Hunt, 1989), is consistent with selective abduction while backward reasoning (goal-based heuristic search - *ibid.*) is consistent with the cycle deduction-induction.

The research relates to the finding that, in solving routine problems in their domains, expert physicians tend to work forward from the available information to hypotheses. On the contrary, intermediate and novice physicians work from a hypothesis regarding the unknown, back to the given information. A strong relationship between diagnostic accuracy and the existence of forward reasoning has been established. (In standard experimental procedure, subjects are shown a written description of a clinical case and each subject is asked to read the clinical test for a specific period of time, after which it is removed. The subjects are asked to write down as much of the text as they can remember, and then to describe the underlying patho-

physiology of the case. Finally, they are asked to provide a diagnosis - Patel and Groen, 1991; see also Groen and Patel, 1988).

All expert physicians with completely accurate diagnoses revealed the use of pure forward reasoning, followed by evaluation in order to confirm and refine the diagnosis by explaining the patient's cues (Patel, Evans and Kaufman, 1989). When experts do not provide complete diagnoses, they use a mixture of forward and backward reasoning, that is, the generation of alternative possibilities (plausible hypotheses), followed by an evaluation phase in which the alternative diagnoses can be discriminated. The difference between accurate and inaccurate diagnoses is the presence of "loose ends". This is also the case for intermediates who do not seem to be able to filter out irrelevant information: this causes the production of loose ends, that is the activation of irrelevant searches. On the contrary, the efficacious use of pure forward reasoning expresses the idea that "a distinguishing trait of experts [...] is a knowledge of what not to do" (Patel and Groen, 1991).

In the case of doctor-patient interactive dialogues, analyzed using linguistic pragmatics methods, these authors argue that "it is expected that physicians initially adopt a data-driven strategy and later shift to a predictive reasoning strategy when they have a working hypothesis; [...] the directionality of reasoning is in forward direction until some loose ends are generated, when the reasoning shifts to the backward direction to account for the loose data" (Patel and Groen, 1991). In this case also, experts arrive at accurate diagnoses because their initial hypotheses are generally accurate, which results in the accurate prediction of subsequent findings.

Forward reasoning remains associated with accurate diagnosis; during this reasoning process scientific biomedical information is not used, whereas during predictive reasoning it is used. On the contrary, residents collect a number of alternative possible diagnoses, and thereby a number of loose ends which produce diagnostic inaccuracy.

In my opinion, the cognitive concept of forward reasoning is consistent with the selective abduction of my model, because both deal with an inference from data to hypotheses. Likewise, as previously mentioned, backward reasoning is consistent with the deduction-induction cycle, because both deal with an inference from hypotheses to data. Nevertheless, in order to avoid any misunderstanding, it is necessary to illustrate various differences:

(1) epistemologically, selective abduction always implies *uncertainty*, although it tends to produce hypotheses that have some chance of turning out to be the best explanation; at this stage it is not known which hypothesis is the best and this type of reasoning does not possess the resources to answer the question; on the contrary, from an empirical cognitive point of view, forward reasoning characterizes an expert's diagnostic accuracy, that is the

diagnostic reasoning that is immediately successful and that establishes the best explanation. The selectivity considered as guessing plausible hypotheses is not relevant, rather forward reasoning seems to be consistent with the philosophical concept of visual abduction described above;

(2) epistemologically, the deduction-induction cycle illustrates inference to the "best" explanation involving some multi-dimensional criteria of evaluation and of the elimination of hypotheses; on the contrary, the empirical cognitive results show that this kind of reasoning is typical of intermediates' diagnostic "inaccuracy" - although they recall better than experts and novices - because of the effect of the failure of forward reasoning, and of the consequent production of unnecessary searches (clearly judged "unnecessary" *post hoc*).

A similar problem was considered and analyzed by Simon (1966), in terms of the classical concepts of "problem-solving" and "selective trial and error search". When Simon observed that an important generalization referring specifically to "the kinds of thinking activities called 'problem-solving'" was that "Problem-solving involves a highly selective 'trial and error' search of solution possibilities" (Simon, 1977, p. 277), he described something analogous to my complete abduction-deduction-induction cycle. He continued:

> Problem-solving searches require trial and error in that they generally do not go directly to the solution without traversing and retracing some blind alleys - sometimes many, sometimes few. When a person solves a problem without any backtracking whatsoever, we are apt to deny that he needed to think at all. We say, "He knew the answer", or "he didn't have to think; he did it by rote" (*ibid.*).

From a psychological point of view, in the first case, when a person is required to think by "trial and error", we have the empirical-cognitive side of my epistemological and normative complete abduction-deduction-induction cycle: reaching the correct solution (i.e., abduction as the best explanation) expresses abduction as involving both the generating and deductive-inductive phases of testing. In the second case, when a person solves a problem directly, the deductive-inductive phase of my complete cycle is missed out and abduction as inference to the best explanation can be accomplished without any testing phase. Moreover, in order to interpret correctly the notion of "best explanation", it is necessary to emphasize, as Simon does, "that human problem solvers and computer programs that simulate them do not search for the 'best' solution, but for the 'best' solution that is 'good

enough' by some criterion" (*ibid.*, pp. 280-1). The best solution always has to satisfy contextual criteria.

3. THE NEED FOR AN EPISTEMOLOGICAL ARCHITECTURE OF MEDICAL KBSs

Since the mid-'80s there has been widespread agreement among cognitive scientists that models of a problem-solving agent should incorporate knowledge about the world (*ontological commitment*) and some sort of an abstract procedure (*inferential commitment*) for interpreting this knowledge in order to construct plans and take action.

Going beyond the level of formalisms and programming tools, the concept of heuristic classification (Clancey, 1985), the distinction between deep and surface models (Chandrasekaran and Mittal, 1982; Steels, 1984), problem-solving methods (McDermott, 1988; Marcus, 1988a), and the idea of generic tasks and task-specific architectures (Chandrasekaran, 1983) took advantage of the increasing epistemological interest of KBSs to elaborate some basic issues to do with reasoning. These fours main approaches have been proposed during the last decade in order to provide a theoretical foundation of KBS (Steels, 1990). They are usually regarded as abstraction paradigms leading to rational reconstructions of KBS. However, their formulation does not represent only a theoretical problem but also a fundamental step in KBS design.

Heuristic classification provided a theoretical framework for KBS stressing features and properties of heuristic reasoning (experiential reasoning in the heuristic classification's lexicon) that a KBS should develop. As an abstraction paradigm, heuristic classification focuses on the inference structure underlying expertise. There are several advantages in an analysis focusing on inference structures: it allows to identify basic components of heuristic pathways by highlighting features that are beyond the domain knowledge and, therefore, it shows similarities and differences across different problem types and application domains. Systems such as NEOMYCIN (Clancey, 1981) and GUIDON (Clancey, 1984) have been developed following the heuristic classification abstraction paradigm.

Another abstraction paradigm - based on the dicotomy *deep-shallow systems* - focuses on the theoretical structure and contents of domain knowledge, instead of inference structure. A system is said to be *deep* when the problem solver works on an explicit symbolic representation of the structure and behavior of the underlying pathophysiological system, while this type of knowledge is only implicitly represented in a shallow system. Following earlier remarks made by Davis (1984), several authors pointed out that the

domain knowledge becomes explicit and more accessible when the inference structure is separated as much as possible from the domain knowledge. However, in order to exploit domain knowledge in a suitable way, a domain-independent inference structure needs to be designed. The explicit and separate representation of domain knowledge and inference structure leads to assume that there is a calculus of thought manipulating knowledge independently from its contents. Thus the definition of the inference structure turns out as a fundamental task also for the deep-shallow system approach.

The paradigm based on the concept of *problem-solving method* is the characterization of the sequence of actions that enables the agent to execute a certain task in a specific domain. A problem-solving method is usually defined by some mechanisms (1) to generate a set of candidate actions and (2) to select among these candidates the action to be executed. For instance, diagnosis (Eshelman, 1988) can be viewed as a process of covering and differentiating (i.e. (1) finding possible diseases covering most symptoms and (2) differentiation between the remaining explanations) and construction (Marcus, 1988a) as a process of proposing and revising solutions (i.e. (1) proposal of a partial solution and (2) revisal of solution by resolving violated constraints). One of the most interesting results of this approach is the claim that each of these problem-solving methods needs to fill domain knowledge into some method-specific roles: diagnosis, for instance, requires knowledge about the relationships linking symptoms to possible diagnosis and additional symptoms to further differentiate. Hence, domain knowledge may be no longer regarded as independent from its concrete use in the problem-solving process, as in deep systems. However, domain knowledge is still explicitly and separately represented and this allows us to gain the advantages of maintainability and systematicness of domain knowledge.

The basic idea underlying *generic tasks* approach is that every real world complex task may be decomposed into simpler subtasks, having input/output relations between them. Each task falls into major classes of tasks, named *generic tasks*. Chandrasekaran identified a small set of generic tasks, such as interpretation, classification, diagnosis, design, and so forth. These generic tasks represent basic elements in the conceptual architecture of a KBS, and they show similarities across application domains. MDX (Chandrasekaran, 1983) was the first medical KBS developed within the generic task framework.

Besides their differences, all these abstraction paradigms, heuristic classification, deep-shallow systems, problem-solving methods, and generic tasks, are not alternative each other. Clancey (1985) showed how heuristic classification is useful in building a deep system as GUIDON; Chandrasekaran (1986) stressed that the generic task approach may allow us to develop a deep system and that the concept of generic task may involve the definition

of a suitable inference structure (Chandrasekaran, 1988). MDX-II provides an example of a deep system developed following the generic task approach. On the other hand, a deep system such as NEOCRIB (Johnson and Keravnou, 1988) has been described in terms of the generic tasks it executes. Finally, domain knowledge is explicitly and separately represented in MOLE (Eshelman, 1988), a tool for developing KBSs based on the cover-and-differentiate method, that is a specific problem-solving method.

These various new approaches prompted the exploitation of the epistemological model (ST-MODEL) described above in order to design the general inferential behavior of NEOANEMIA (Stefanelli, et al., 1988; cf. the following section). Moreover the model ought not to be judged by how faithfully it represents human processing unless that is its very point (Glymour, 1992, p. 365), but this does not mean that the way people seem to reason is not a matter of consideration when designing KBSs (Evans and Gadd, 1992).

The epistemological model needed to be made more complex with an *ontological level* dealing with the entities and relationships comprising the domain knowledge of the KBS. Different ontologies express diagnosis, therapy planning and monitoring, but to solve problems, the three tasks can be carried out by a single inference procedure in terms of abduction, deduction and induction (see section 2, above). The KBS ontology that adequately and "deeply" represents knowledge, as it is organized in scientific medical theories (causal or taxonomic) (Kuipers, 1987; Milne, 1987; Simon, 1985), goes beyond first generation "shallow" KBSs that only mapped knowledge into pragmatic constructs derived from human experts - in the latter case the ontology was compiled in conjunction with the inference procedure, thereby becoming implicit - (Chandrasekaran, 1983). In this sense the new architectures combine a more principled knowledge of the domain with the simple heuristic knowledge that was the main type of knowledge exploited in first generation KBSs.

The need for representing various kinds of medical knowledge in a working KBS has specific consequences at the level of implementation. Ideally, it can be concluded that the choice of a convenient methodological medium originating from various disciplines (AI, mathematics, statistics, probability theory, decision theory, logic, and so on) for representing ontologically real medical knowledge (and to create the inference process) must be connected to the complexion of that knowledge and to the cognitive task at hand. A single method or formalism is not able to interpret efficiently all types of knowledge. Indeed the ontology that is embodied in intelligent computation is typically done with a certain inferential purpose in mind, and a good inference under computational and environmental constraints often requires various methods of representation, many formalisms and a mixture of the two. Exploiting these methods may actually improve performance and

allow physician to explore the implications of certain observations or hypotheses to predict the effects of actions or the effects of future data on his/her behavior.

4. NEOANEMIA

The architecture of NEOANEMIA[3] (for further details on how the epistemological model works in the context of this KBS see Lanzola, et al., 1990), a working diagnostic system to manage disorders causing anemia, follows the epistemological abduction-deduction-induction model (ST-MODEL), as a shared framework permitting the knowledge engineer and the physician to cooperate in the development of the system thus overcoming the esoteric character of the underlying implementation.

NEOANEMIA derived from a previous KBS called ANEMIA (Quaglini et al., 1986) which was proved to perform quite satisfactorily. According to the epistemological model of the diagnostic task, NEOANEMIA starts by selecting, via *abduction*, general diagnostic categories from clinical evidences obtained through an abstraction of routine hematologic findings. All NEOANEMIA's entities, i.e. patient's data and diagnostic entities, are organized into taxonomic structures. The system proceeds by exploring and possibly refining among the previously abduced hypotheses. During its abductive inference, NEOANEMIA exploits a knowledge base built up using production rules and generates a diagnostic space containing all the abduced diagnostic hypotheses. Then, diagnostic space structuring takes place: NEOANEMIA mainly looks for compatibilities or possible associations, thus grouping or differentiating among the diseases included in the diagnostic space. In such a way, composite hypotheses can be assembled.

Once abduction and diagnostic space structuring have been executed, NEOANEMIA will *deduce* expected manifestations from the hypotheses included into the diagnostic space. While in the hypothesis generation phase, NEOANEMIA exploits compiled *heuristic pathways* specified by the expert, a separate and explicit representation of *taxonomic* and *causal ontology* is used in the testing phase. They have been represented using respectively a simple two layer network (i.e. representing which clinical evidences may be expected in presence of each disease) and QSIM (Kuipers, 1986, 1987; Ironi, et al., 1992) for representing available knowledge on pathophysiological system dynamics. QSIM requires that the basic etiological mechanisms of a disease are represented as perturbations of the initial conditions of a qualitative model of iron metabolism and expected manifestations are then found

[3] I collaborated with the development of this computational program at the Department of Computer Science of the University of Pavia, Italy.

by analyzing the new reached steady state conditions. QSIM models provide a deeper understanding of portions of knowledge expressed by the two level network. They also predict the effect of multiple diseases acting together.

The *inductive* phase is then carried out by comparing expected with observed findings, when available, or by planning what to do next. Further actions aim either to corroborate candidate hypotheses or to discriminate between competing hypotheses, or to rule-out hypotheses. The selected action depends on the contents of the structured diagnostic space and the strategy to make the decision is represented in the control knowledge by meta-rules. In order to represent explicitly the taxonomic ontology exploited in the deductive phase, we enhanced the frame-based language we used with an ATMS (De Kleer, 1986) (cf. chapter 2, section 2), allowing us to handle the non-monotonic character of the deductive phase: the hypothesis hierarchy is mapped into a hierarchy of environments each representing different possible states of the patient as suggested by abduced hypotheses.

Computational problems arising in the implementation of the epistemological model of therapy planning have been taken under careful consideration. To this aim a KBS called Therapy Advisor (TA) has been developed (Quaglini et al., 1989). The task of TA is to plan an adequate therapy depending on the etiopathologic mechanisms causing a patient's anemia as diagnosed by NEOANEMIA.

According to the epistemological model of the therapy planning task, TA starts from deriving a restricted set of critical aspects of a patient's condition which might be eliminated adopting a suitable therapy. These represent abstractions which have been categorically derived from patient's data and formulated diagnosis in order to abduce a set of therapies potentially useful to handle those problems. For most of them a well agreed-on therapeutic plan can be proposed. In such cases, therapy planning does not involve basic strategic choices, but rather consists in fixing details such as dosage and route of administration in order to meet patient's conditions. Thus, production rules have been judged as the most appropriate formalism for representing knowledge. The remaining therapeutic problems involve trade-offs between conflicting goals, in presence of uncertainty about the therapy effects. In presence of uncertainty and trade-off, the conclusion about the recommended therapeutic action may depend on a large number of patient's data. They are usually called "defeaters". For example, the decision to remove the spleen (splenectomy) in a myelofibrotic patient in order to reduce his/her transfusional need interacts with the "platelet count", since in presence of a high platelet count there is a significant risk of splenectomy-induced thrombosis.

The presence of several defeaters would burden the rules with a large number of possible exceptions whose pathophysiological meaning would be

difficult to be understood by a non-expert user. Moreover, it is essential for building an advanced TA to represent explicitly and separately the medical knowledge the system uses to predict the effects of a therapy, possibly through a model of the underlying pathophysiological system, and the preferences, elicited from either the user or the patient.

These reasons led to the choice of adopting "influence diagrams" (IDs) for representing the most challenging therapeutic decision problems. IDs are directed acyclic graphs with three types of nodes: decision, chance, and utility nodes. Decision nodes represent choices available to the decision-maker. Chance nodes represent uncertain quantities, that may either characterize the patient's condition at the decision time or therapy outcomes. Finally, the utility node embeds the utilities assigned to possible configurations of values of chance nodes.

TA's knowledge base contains some taxonomies of entities represented by means of frames: patient's findings and diagnostic entities are shared with NEOANEMIA, while therapeutic problems and adoptable therapies are specific portions of knowledge for therapy planning. Starting from observed patient's data and diagnosis formulated by NEOANEMIA, TA selects the relevant therapeutic problems in the patient at hand. To this aim TA exploits different kinds of pathophysiological concepts: some of these are abstracted by NEOANEMIA from patient's data (for example, "severe anemia", "augmented serum bilirubin", etc.), some others represent general categories of diseases included by NEOANEMIA into its diagnostic space (for example, "aregenerative anemias", "hemolytic anemias", etc.) able to define in a synthetic way the behavior of the erythropoietic system, while the remaining concepts are abstracted by TA from patient's data and are needed for selecting the most suitable therapy.

To summarize, NEOANEMIA employs different representation formalisms and methods to achieve selective abduction and the deduction-induction cycle. Moreover the inference procedure (control knowledge) and the ontology (domain knowledge) are explicitly and separately represented. While in the hypothesis generation phase NEOANEMIA exploits compiled heuristic pathways specified by an expert, a separate and explicit representation of causal and taxonomic ontology is used in the deduction-induction phase.

It is important to remember some of the considerations of section 2. In a diagnostic KBS selective abduction does not always imply uncertainty: sometimes in medical KBSs the selective abduction phase provides the best hypothesis immediately, that is the best explanation, because the selection has been very successful, to a certain degree simulating the efficacy of experts' forward reasoning. In this case the so-called evaluation-testing phase (deduction-induction) only provides the opportunity for an "explanation" of the abduced hypothesis, without performing any unnecessary discriminating

movement. Thus the explanation can exploit the basic medical knowledge (causal or taxonomic) in KBS ontological models to make, for instance, communication between physicians or the processes of teaching and learning easier. Moreover, a medical KBS can make the transition to the deduction-induction phase, in order to exploit this cycle and to reduce previously abduced hypotheses, so as to reach the correct diagnostic conclusion. This does not reflect, as shown by experimental sciences on expertise (see section 2, above), the weakness (i.e. inaccuracy) of low-level diagnostic performances (backward reasoning): instead the cycle reflects the application of a powerful knowledge base. Diagnostic KBS behavior in this case is exactly the same as for the epistemological model: it expresses inference to the best explanation involving some multi-dimensional criteria of evaluation-elimination of hypotheses; the criteria are produced at the computational level by suitable methods of representing ontological and inferential commitments of the deduction-induction phase. As stated above, medical KBSs of this kind ought not be judged by how faithfully they represent human processing because this is not the precise point. These KBSs may be considered as mental prostheses (I would add "rational and objective") that assist physicians with different skills and expertise in the management of patients. Just as telescopes are designed to extend the sensory capacity of humans, KBSs are designed to extend their cognitive ability.

Finally it is important to note that there is a fruitful exchange between AI and epistemology. On the one hand, as described above, there is the need for an abductive epistemological architecture of medical KBSs, on the other hand it is important to note that there are well-known AI systems that perform explicitly epistemological tasks (cf. chapter 2, section 5). The increase in knowledge provided by this intellectual interaction is manifest.

If the pure philosophical and cognitive task is to state correct rules of reasoning in an objective way, the use of computer modeling may be a rare tool in these investigations because of its rational correctness. This cooperation should prove very fruitful from an educational perspective too: reciprocally clarifying both philosophical and AI theories of reasoning will provide new and very interesting didactic tools.

5. BASIC SCIENCE REASONING AND CLINICAL REASONING INTERTWINED

From an epistemological point of view (Schaffner, 1986) biomedical sciences can be considered as a set of *partially overlapping models* (sometimes built at the cross-roads of several disciplines) of semi-independent phenomena dealing with prototypical cases. The role of generalizations in biomedi-

cal sciences is to use explicitly *exemplars* - exemplars are identified by Kuhn (1970) as the accepted, prototypical problems that can be encountered both when learning a discipline and in discussion of its contemporary research - and to capture causal relations between them, whereas the role of generalizations in the physical sciences is to give abstract laws relating to several exemplars. In the clinical biomedical sciences exemplars also concern individual's abnormalities:

> This implies that an important, perhaps implicit, component of medical theory involves models of normative biomedical behavior. Since that, too, may be based on sets of exemplars, we see the possibility that clinical medicine, if a scientific theory, is a theory based on models of models - clearly not a straightforward product of axioms of biology (Patel, Evans and Groen, 1989a, p. 56).

The Kunhian concept of exemplar refers to the field of growth of scientific knowledge and in this sense is related to the "anti-theoretical" emphasis on problem-solving performance:

> Philosophers of science have not ordinarily discussed the problem encountered by a student in laboratories or in science texts [...] at the start and for some time more, doing problems is learning consequential things about nature. In the absence of such exemplars, the laws and theories he has previously learned would have little empirical content (Kuhn, 1970, pp. 187-8).

In cognitive science this (and similar) type of postpositivistic objection to the formalistic excess of the neopositivistic tradition in philosophy of science is exploited to stress the relevance of the distinction between theories and their domains of application. This objection is exploited to emphasize the difference between established bodies of scientific knowledge and their processes of discovery and/or application and, in medical knowledge, between clinical reasoning (situated, concerned with attributes of people) and basic science reasoning (unsituated, concerned with attributes of entities such as organs, bacteria, or viruses).

There have been many experimental studies (see section 2, above) in cognitive psychology to elucidate the precise role of *basic science* in medical problem-solving in order to determine: 1) to what extent basic science and *clinical knowledge* are complementary; 2) what basic science contributes to medical problem-solving; and 3) whether basic science knowledge contributes to medical expertise (Patel, Evans and Groen, 1989a); see also (Groen and Patel, 1988; Patel, Evans and Kaufman, 1990). The distinction between basic medical science (and reasoning), and clinical science (and reasoning) is also included in the general problem of *medical education* (see the following

section). The AI ways of exploiting basic science resources in ontological levels involved in the deduction-induction cycle of second generation medical KBSs are described above. This is the case with NEOAMEMIA, but applies equally to earlier medical KBSs, although differently, such as CASNET, CADUCEUS (Pople, 1985) and ABEL (Patil, 1981).

The aim here is to outline some basic philosophical issues that may help to clarify the problem of medical education, at least from a theoretical point of view. We have described in chapter 1 that the problem of "teaching" science is a very old topic of philosophical reflection. Plato's *Meno* is a dialogue about whether virtue can be taught (Turner, 1989): the problem is related to the *Meno* paradox, stated by Plato in the dialogue and discussed by Simon (1976), and to the issue of "tacit knowledge" which was introduced by Polanyi (1966).

The story of *Meno*'s slave can be looked at from the point of view of an epistemological argument about the paradoxical concept of "problem-solving". Plato's solution of this epistemological impasse is the very classic philosophical scenario of the doctrine of reminiscence: Socrates' teaching is in reality leading the slave to discover the knowledge he already possesses in his spirit.

The slave boy in the dialogue is brought in to make a related point:

> Socrates establishes a) that the boy cannot correctly answer the question ("cannot tell", in Polanyi's language), of how much larger the sides of a square with double the area of another square will be, and b) that the boy thinks he knows that if a square has twice the area the sides will also be doubled. He then leads the boy through a series of inferences, each of which the boy could "tell", at least could assent in response to Socrates' "questions" formulating those influential steps, and that he could correctly multiply and add when asked (Turner, 1989, p. 85.).

These queries lead the boy to the correct answer. As illustrated in chapter 1, Polanyi thinks the *Meno* story shows that if all knowledge is explicit, i.e., capable of being clearly stated, then we cannot know a problem or look for its solution. It also shows that if problems nevertheless exist, and discoveries can be made by solving them, we can know things that we cannot express: "to search for the solution of a problem is an absurdity; for either you know what you are looking for, and then there is no problem; or you do not know what you are looking for, and then you cannot expect to find anything" (Polanyi, 1966, p. 22).

In his turn, Simon provides a computational solution of the paradox in modern problem-solving terms: "our ability to know what we are looking for does *not* depend upon our having an effective procedure for finding it: we need only an effective procedure for testing candidates" (Simon, 1977, p.

339). If it is possible to have an effective procedure for testing, and a procedure for generating candidates, we will have a "problem", i.e. an unsolved problem, where we nevertheless "know what we are looking for" without actually possessing it.

I recalled the *Meno*'s story again in order to illustrate a prototypical "cognitive" story, from philosophical to knowledge engineering outcomes. Socrates teaches the slave some geometric issues in a problem-oriented fashion, not a theorematic one (but this is before Euclid's *Elements*). He shows the slave some inferential routines and subroutines (for recognizing numerical inconsistency or for calculating area, for instance) for generating and testing (in Simon's terms) that enable him to self-program (or "learn") and solve the problem, thus coming to know new geometric notions.

These observations delineate the centrality of the concept of problem-solving in teaching and learning. There is no longer room for a philosophical doctrine of reminiscence. New developments consist of benefiting from recent rational clarifications of problem-solving and problem-oriented knowledge due to artificial intelligence and cognitive science. Thus the philosophical story above introduces the main methodological issues in medical education. In medical training the following ideas need to be emphasized and added to conventional curricula:

1. the need for an epistemological and logical (didactic) awareness of the main methodological topics (for instance, abduction) incorporated in reasoning for diagnosis, therapy planning and monitoring;
2. the relevance of problem-oriented teaching and learning, as different from conventional basic science-centred education, and its relations and interaction in education itself and in reasoning performances;
3. the role of KBSs, tutoring systems (Clancey, 1986; Kunstaetter, 1986) and other technological products in allowing, for instance, students to browse ontologies that express stored basic medical knowledge and to see reasoning processes displayed separately and explicitly during computational problem-solving.

To exploit the educational consequences of the previous analysis I will try to answer some questions: What is the role of problem-solving in teaching and learning, as different from conventional basic science-centred education? Is it relevant, in medical education, an epistemological and logical awareness of the main methodological topics?

6. COGNITIVE MODELS AND MEDICAL EDUCATION

It is interesting that *conventional curricula* (CC) (where basic science courses are taught before the clinical training) and *problem-based learning curricula* (PBL) (where basic science is taught in the context of clinical problems and general heuristics are specifically taught) lead students, when they generate explanations, to develop respectively selective abductions (forward reasoning) or to perform the whole abduction-deduction-induction cycle using relevant biomedical information (backward reasoning). The results of this cognitive research can be found in Patel, Groen and Norman, 1993; see also Patel, Evans and Groen, 1989b.

Educational transfer occurs when knowledge acquired in a specific context or for one purpose is used in a different context or for a different purpose. Many kinds of transfer involve the application of knowledge in performing basic cognitive tasks; these tasks are performed in conditions that differ from those under which the pertinent knowledge was acquired.

Following Patel, Evans and Groen (1989a), we can affirm that clinical medicine and the biomedical sciences constitute two distinct and not entirely compatible worlds, with separate ways of reasoning and quite different ways of structuring knowledge. Clinical knowledge is based on a complex taxonomy which relates diseases and symptoms to the underlying pathology. On the contrary, the biomedical sciences are based on very general principles which define chains of causal mechanisms. Thus, learning how a set of symptoms relate to a diagnosis may be very different from learning what causes an illness. In both cases learning is not only epistemologically but also situationally directed. Clinical training is built on situations involving hospital and patients. The biomedical sciences are built on laboratory situations. However, the latter are learnt primarily from textbooks whereas the former is based on living encounters. Clinical training is highly situated, whereas biomedical instruction is highly unsituated. Hence, it may be better to teach basic science separately, so that the appropriate knowledge can be activated in spontaneous problem-solving situations.

The persistence of scientific errors in PBL students' explanations indicates that there are serious problems as regards the specific implementation of the PBL curriculum. If students merely gather experiential knowledge, they run the risk of building a huge situation-knowledge base that is not connected with theoretical, general biomedical knowledge. Moreover, it is well-known that the deliberate use of a rigid epistemological mechanism may also interfere with the proper richness of an expert's forward reasoning.

The PBL curriculum specifically involves the specific teaching of hypothetico-deductive reasoning, that is, the reasoning illustrated by the whole

cycle of my epistemological model. Consequently, students learn a systematic process of thinking - yet do they end up with an epistemological awareness of the main methodological topics involved?

In these students it has been empirically noted the systematic use of clinical information and the tendency to elaborate extensively. Nevertheless, in CC students we have not observed a systematic mode of reasoning: the mode of reasoning most often employed is "forward", together with the tendency to explain cases on the basis of a single diagnosis rather than an extensive list of differential diagnoses. It has been also observed a lack of scientific background in many PBL students; on the contrary, the surprising degree of structure used by CC students in their explanations may be due to their superior scientific background. The active application of knowledge ("basic science") in understanding clinical texts and reasoning about problems will lead students to construct stronger relations between concepts, increased coherence of knowledge networks and an increased number of interrelations between concepts. In addition, a process of knowledge restructuring takes place.

Boshuizen and Schmidt (1992) have shown that the repeated application of biomedical knowledge in clinical reasoning at the earlier stages of expertise development leads to the subsumption of lower-level detailed propositions under higher-level, often clinical, propositions. This is the well-known phenomenon they termed *encapsulation*: the result consists of easily accessible and flexible knowledge structures with very short research paths. Biomedical knowledge plays a tacit role since it is encapsulated in clinical knowledge. Our findings suggest a tacit role for biomedical knowledge in expert clinical reasoning. This tacit role contradict the conviction (Patel, Evans and Groen, 1989a) that biomedical and clinical knowledge essentially represent two different worlds.

Empirical results have in turn shown that an expert facing familiar problems applies increasingly less detailed medical knowledge than experts facing unfamiliar ones. Knowledge applied in such routine cases has already become encapsulated. Furthermore results from Boshuizen and Schmidt state that the detailed biomedical knowledge encapsulated under high-order propositions remains available and can be retrieved whenever necessary, for instance in explanations and communications. For example, when asked to explain the direct connection to be made between drug abuse and endocarditis Patel and Groen's subjects would probably easily expand that higher order proposition to a chain of at least five other propositions.

Boshuizen and Schmidt's model is complicated by the notion of *illness script* (Feltovich and Barrows, 1984). The illness script is caused by that process that takes place when a student acquires proficiency in clinical reasoning, when he or she is exposed to real patient. All illness scripts are as-

sumed to develop as a result of extended practical experience: illness scripts will tend to become richer, more refined, and better turned to practice, while causal, biomedical knowledge which they incorporate becomes further encapsulated as a function of the amount of actual experience with a certain (class of) disease(s).

We already said that, since the mid-'80s, there has been widespread agreement among AI scientists that models of a problem-solving agent in a KBS should incorporate knowledge about the world (ontological commitment) and some sort of an abstract procedure (inferential commitment) for interpreting this knowledge in order to construct plans and take action.

As already illustrated in sections 3 and 4 above, there are many AI ways of exploiting basic science resources in ontological levels involved in the deduction-induction cycle of second generation medical KBSs. This is the case with NEOAMEMIA (Lanzola, et al., 1990) but also with earlier medical KBSs, such as CASNET, CADUCEUS (Pople, 1985) and ABEL (Patil, 1981). The KBS ontology that adequately and "deeply" represents knowledge, as it is organized in scientific medical theories (causal or taxonomic) goes beyond first generation "shallow" KBSs that only mapped knowledge into pragmatic constructs derived from human experts - in the latter case the ontology was compiled in conjunction with the inference procedure, thereby becoming implicit. The new architectures combine a more principled knowledge of the domain with the simple heuristic knowledge that was the main type of knowledge exploited in first generation KBSs. We have to remember that, while in the hypothesis generation phase NEOANEMIA exploits compiled heuristic pathways specified by an expert, a separate and explicit representation of causal and taxonomic ontology is used in the deduction-induction phase.

7. THE CENTRALITY OF ABDUCTION

The concept of abduction is philosophically very powerful: I have sought to show in the previous pages its efficacy in unifying many intellectual areas devoted to the clarification of problem-solving processes and medical reasoning. In my opinion these abductive schemes may form a forceful language capable of establishing a relatively solid and objective framework that increases the intelligibility of many cognitive phenomena.

Peirce's insight about the "inferential" virtues of abduction has been proved very far-sighted. AI, logical and cognitive studies of problem-solving processes have guaranteed the philosophical centrality of abduction in present-day cultural, scientific and technological developments. Simon's observation that abduction "is the main subject of the theory of problem-solving"

has been ratified. This centrality attracts a cluster of related topics, from logic of discovery to evaluation skills, from nonmonotonic logic to medical, spatial and temporal reasoning, from AI systems to the detection outlook in narrative contexts (Eco and Sebeok, 1983): I will consider some of them in the following chapters.

I have unified medical reasoning by the notion of abduction: this kind of reasoning explains and executes the three generic tasks of diagnosis, therapy planning, and monitoring, correctly establishing the level of evaluation procedures and ontological medical complexity.

The relevance of abduction ensures it a prominent role in methodological aspects of medical education and practice. Moreover I have tried to show that the idea of abductive reasoning might be a flexible epistemological interface between other related notions (induction and deduction, best explanation, perception, forward and backward reasoning, defeasibility, discovery, and so on) all of which are involved in medical reasoning and in medical education but, at the same time, are of great theoretical interest in general.

I have tried to clarify the role of problem-solving in medical education, as different from conventional basic science-centred education. Although an epistemological and logical awareness of the main methodological topics concerning medical science and reasoning would be important from a general theoretical point of view, it does not improve the physician's performance. Anyway, does the PBL curriculum actually involve the specific teaching of hypothetico-deductive reasoning, that is the reasoning illustrated by the whole cycle of my epistemological model? Do these students end up with an epistemological awareness of the main methodological topics involved?

The empirical results show that backward reasoning is typically of intermediates' diagnostic "inaccuracy" - although they recall better than experts and novices - because of the effect of the failure of forward reasoning, and of the consequent production of unnecessary searches. Moreover, the persistence of scientific errors in PBL students' explanations indicates that there are serious problems as regards the specific implementation of the PBL curriculum. If students merely gather experiential knowledge, they run the risk of building a huge situation-knowledge base that is not connected with theoretical, general biomedical knowledge: there is a lack of scientific background.

According to Boshuizen and Schmidt biomedical knowledge plays a tacit role since it is encapsulated in clinical knowledge as a function of the amount of actual experience with a certain class of diseases. This tacit role would contradict Patel and Groen's conviction that biomedical and clinical knowledge essentially represent two different world.

Chapter 5

Visual and Temporal Abduction

1. VISUAL ABDUCTION

1.1 Visual imagery

The general objective is to consider how the use of visual mental imagery in thinking may be relevant to hypothesis generation and scientific discovery. To this end I treat imagery as a problem-solving paradigm in artificial intelligence and illustrate some related cognitive models. In this research area the term "image" refers to an internal representation used by humans to retrieve information from memory. Many psychological and physiological studies have been carried out to describe the multiple functions of mental imagery processes: there exists a visual memory (Paivio, 1975) that is superior in recall; humans typically use mental imagery for spatial reasoning (Farah, 1988); images can be rebuilt in creative ways (Finke, and Slayton, 1988); they preserve the spatial relationships, relative sizes, and relative distances of real physical objects (Kosslyn, 1980); for a more complete list, see Tye (1991).

Kosslyn introduces visual cognition as follows:

> Many people report that they often think by visualizing objects and events [...] we will explore the nature of visual cognition, which is the use of visual mental imagery in thinking. Visual mental imagery is accompanied by the experience of seeing, even though the object or event is not actually being viewed. To get an idea of what we mean by visual mental imagery, try to answer the following questions: [...] How many

windows are there in your living room? If an uppercase version of the letter *n* were rotated 90° clockwise, would it be another letter? (Kosslyn and Koenig, 1992, p. 128).

We can build visual images on the basis of visual memories but we can also use the recalled visual image to form a new image we have never actually seen. Certainly, imagery is used in everyday life, as illustrated by the previous simple answers, nevertheless imagery has to be considered as a major medium of thought, as a mechanism of thinking relevant to hypothesis generation. Some hypotheses naturally take a pictorial form: the hypothesis that the earth has a molten core might be better represented by a picture that shows solid material surrounding the core.

There has been little research on the possibility of visual imagery representations of hypotheses, despite abundant reports (e.g. Einstein and Faraday) that imaging is crucial to scientific discovery, but also in creative literary and artistic realizations (Koestler, 1964; Shepard, 1988, 1990). Einstein described having imagined the consequences of traveling at the speed of light, which led him to the discovery of the theory of special relativity. Faraday claimed to have visualized lines of force that emanated from elecrical and magnetic sources, leading to the modern conception of electromagnetic fields (cf. chapter 3, section 1, this book). Recently, also the physicist Feynman claimed to have resorted to visual images in speculating about interactions among elementary particles to develop the so-called "Feynman diagrams". Moreover, it is well-known that the German chemist Kekulé, used spontaneous imagery to discover the structure of benzene; Watson and Crick have reported the use of mental imagery in the interpretation of diffraction data and in the determination of the structure of the DNA molecule (Holton, 1972; Miller, 1984, 1989; Magnani, Civita, and Previde Massara, 1994; Nersessian, 1995a; Shepard, 1988, 1990; Thagard, Gochfeld, and Hardy, 1992; Tweney, 1989).

Thus, after illustrating the computational imagery representation scheme proposed by Glasgow and Papadias (1992), together with certain cognitive results, I will explore whether a kind of hybrid imagery/linguistic representation architecture can be improved and used to model image-based hypothesis generation, that is to delineate the first cognitive and computational features of what I call *visual abduction* (a type of model-based abduction, cf. chapter 2).

1.2 Knowledge representation scheme

The central theme of the recent imagery debate in the cognitive science has concerned the problem of representation.

How can we represent images? Are mental images represented depictively in a picture or like sentences of descriptions in a syntactic language?

According to Kosslyn's *depictionist* or *pictorialist* view (Kosslyn, 1983), mental images are quasi-pictures represented in a specific medium called the visual buffer in the mind. Kosslyn builds the analogy that visual information stored in memory can be displayed as an image on a CRT screen. Kosslyn's model of mental imagery proposes three classes of processes that manage images in the visual buffer: the *generation process* forms an image exploiting visual information stored in long-term memory, the *transformation process* (for example, rotation, translation, reduction in size, etc.) modifies the depictive image or views it from different perspectives, the *inspection* process explores patterns of cells to retrieve information such as shape and spatial configuration. According to Pylyshyn's *descriptionist* view (1981, 1984) mental imagery can be explained by the tacit knowledge used by humans when they simulate events rather than by a pictorialist view related to the presence of a distinctive mental image processor.

As shown by certain experimental results, Hinton proposes that imagery involves viewer-centered information appended to object-centered structural descriptions of images (1979). Finke (1989) proposes five unifying principles of mental imagery to summarize some of the properties of imagery that a computational model should represent. Finally, Tye has recently suggested a theory, related to some of the aspects of Marr and Nishihara's theory of visual perception (Marr and Nishihsara, 1978), in which mental images are represented as interpreted, symbol-filled arrays (Tye, 1991).

In Glasgow and Papadias computational model (1992, 1993) the imagery processes are simplified by a depictive knowledge representation scheme that involves inferencing techniques related to generation, inspection and transformation of the representation. According to Kosslyn's cognitive model, the knowledge representation scheme of mental imagery is composed of two different levels of reasoning, *visual* and *spatial*, the former concerned with what an image looks like, the latter depending on where an object is located relative to other objects. The different representations of these ways of reasoning exist at the level of working-memory and are generated from a *descriptive representation* of an image stored in long-term memory in a hierarchical organization. Information is accessed from long-term memory by means of standard retrieval, procedural attachment and inheritance techniques (see Figure 1).

The spatial representation outlines the image constituents in a multidimensional, symbolic array (More, 1981), that maintains spatial and topological properties and avoids distracting details, in accordance with the features of the image domain and the questions involved; moreover, the arrays are nested, to express multiple levels of the structural hierarchy of images. Of course the array can be interpreted in different ways depending on the application (Figure 2).

The spatial representation is processed using primitive functions that *transform* and *inspect* the symbolic array. The "uninterpreted" visual representation depicts the space occupied by an image as an *occupancy array*. The related primitive functions concern operations for manipulating and retrieving geometric information, such as volume, shape and relative distance (rotate, translate and zoom, which change the orientation, location or size of a visual image representation) (Figure 1). In a nested symbolic array each symbol in the array corresponds to a frame in long-term memory: we can extract a subimage and express it by its symbolic array parts.

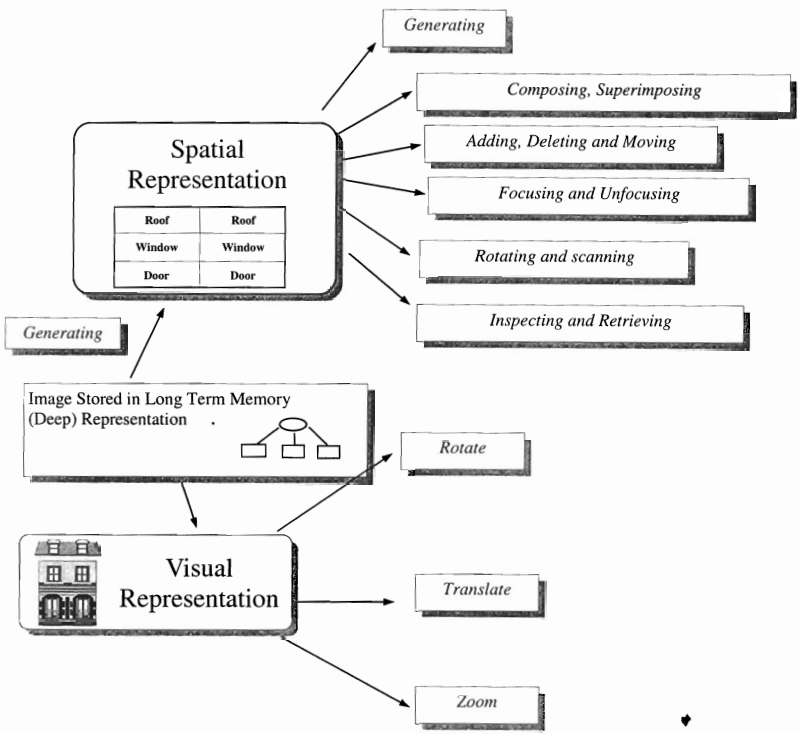

Figure 1. The multi-level knowledge representation scheme.

According to Kosslyn's theory of imagery, the separate long-term memory contains visual and spatial information as descriptions (1980). These descriptions are considered in Glasgow and Papadias computational model as a hierarchical network model for semantic memory: the implementation is done using a frame system where each frame contains notable information concerning a particular image or a class of images. AKO (a kind of), used for property inheritance, and PARTS, to indicate the structural decomposi-

tion of complex images, are the two kinds of image hierarchies in this scheme (Glasgow, 1993).

For example, we can generate as a symbolic array the following descriptive representation of spatial relations (Glasgow, 1993):

The spoon is to the left of the knife.
The plate is to the right of the knife.
The fork is in front of the spoon.
The cup is in front of the knife.

The symbolic array could be:

spoon	knife	plate
fork	cup	

while the corresponding representation in long-term memory could be composed by the following frame structure:

FRAMENAME: place-setting
AKO: image
PARTS/LOCATION: spoon (1,1) knife (1,2) plate (2,1) fork (2,1) cup (2,2)

				Sweden	
Wales / Scotland / England			Denmark		
		Holland	Germany	Germany	
		Belgium			
		France	Italy	Yugoslavia	Yugoslavia
Portugal	Spain		Italy		Greece

Embedded symbolic array representation

Figure 2. Embedded symbolic array representation. (Reproduced from Glasgow and Papadias, 1992, copyright, Cognitive Science Society Incorporated, used by permission).

Moreover, a multi-dimensional array is able to depict spatial (e.g. left-of, north-of) and topological (e.g. adjacent-to, inside) relations among image parts and can also denote non-spatial dimensions such as speed. Sometimes, a part occupies more than one element in the array, to capture all the spatial relationships involved. Hence, an image frame is related to an individual image or to a class of images sharing the same features. Images can be represented in a viewer- or an object-centered frame of reference: unlike Kosslyn's approach, in Glasgow and Papadias computational model (Glasgow, 1993) image representations are typically three-dimensional and object-centered (viewer-independent). Finally, in a nested symbolic array each symbol in the array corresponds to a frame in long-term memory: we can extract a subimage and express it by its symbolic array parts.

Glasgow discusses the concept of *interpretation* of a symbolic array as follows:

> For a given domain with associated spatial relations *R*, the interpretation of a symbolic array representation depends on a set of primitive functions defined in array theory. For example, in a map domain with corresponding relation *R* = {*north-of*, *south-of*, *east-of*, *west-of*, *borders-on*}, we can define a domain function that takes as parameters a symbolic array representation of a map containing several countries and a symbol for a country in the map and returns a list of all the countries that border-on the specified country. Similarly, in a block world, functions might be defined for a set R = {*on-top-of*, *below*, *left-of*, *right-of*, *in-between*}. Symbolic arrays can also be used to denote more complex spatial and topological arrangements between parts of an image, such as the concept of inside/outside or the molecular concept of a 6-member ring cycle. In addition, arrays can be used to express other comparative dimensions such as time, size, speed or age (Glasgow, 1993).

The functions implemented express simple operations and reasoning techniques based on the transformation and inspection of image representations: 1) generating spatial representations from long-term memory; 2) composing and superimposing image representations; 3) adding, deleting and moving parts of an image; 4) focusing and unfocusing attention on parts of an image; 5) rotating and scanning a spatial array; and 6) inspecting and retrieving spatial relations from an image representation. Some of these functions have been implemented into the functional programming language NIAL (Nested Interactive Array Language) (Jenkins, et al., 1986) (Figure 1).

Glasgow and Papadias scheme seems to have some of the features that Johnson-Laird (1983) has advocated as desirable characteristics of mental representations: it achieves conclusions without laborious chains of inferences. It is also consistent with Johnson-Laird's multiple model of imagery

according to which we have to consider a propositional representation, a mental model (as the structural - spatial - layout), and a visual image corresponding to the perceptual viewer-centered representation.

Glasgow discusses Johnson-Laird's influence as follows:

> The motivation for the proposed multiple representations was not expressibility; no new information is gained from the visual and spatial representations. Rather, the distinction is analogous to the issue of high-level programming languages: Why do we have languages with abstract data types such as arrays when all computation can be carried out by manipulating strings of symbols on a Turing machine? Obviously, such abstract structures exist to aid in the process of programming. In the development of our representation scheme for computational imagery, we were motivated by the same distinction. This scheme includes a descriptive representation, corresponding to Johnson-Laird's propositional representation, and two depictive representations, corresponding to the mental model and the visual image. The depictive representations do not necessarily increase the expressive power of the scheme; they do, however, provide a high-level representation for the spatial and visual properties of an image and imply greater programming ease and efficiency when implementing the processes of imagery (1993).

Finally, I should remember that the proposed multiple representation for mental imagery can account for the conflicting experimental data that cannot be explained by either a descriptive, visual or spatial representation.

An immense amount of results about imagery, but also visual, spatial, and diagrammatic reasoning, characterized the last decade of research in logic, cognitive science, and artificial intelligence[1]. In addition to the symbolic array theory described in this section, which may be too restrictive for certain purposes, there are many other cognitive models of imagery (and visual) representations and procedures, that may have particular advantages: in terms of shape grammars (Leyton, 1989, 1992), graph grammars (Thagard and Shelley, 1997), chunks (Anderson and Libiere, 1998), and records of neural activation that occurred during perception (Barsalou, 1999).

Some computational programs which can create analogies with imagistic information are of direct interest for the problem of visual abduction. Letter-Spirit (McGraw and Hofstadter, 1993) takes a stylized letter as input and

[1] The Journal *Computational Intelligence* devoted a special issue to the so-called "imagery debate" (9, 1993). Moreover, the AAAI Society organized the "Spatial and Temporal Reasoning" Workshop in 1994, and a spring Symposium on "Cognitive and Computational Models of Spatial Representation", in 1996, not to mention the many Journals dealing with visual imagery, diagrammatic reasoning, and spatial ability in the area of cognitive psychology.

outputs an entire font that is of the same style. The Structure Mapping Engine (SME) (Falkenhainer, Forbus, and Gentner, 1990), is a kind of agent that finds the best mapping of elements between two given systems: the system has been applied to imagistic knowledge in a system called MAGI, that uses SME to find examples of symmetry and repetition in a single image, detecting similarities in the structure of parts of the images. The VAMP systems are analogical mappers (Thagard, Gochfeld, and Hardy, 1992). VAMP.1 uses a hierachically organized symbol/pixel representation, that detects overlapping pixels in couples of images. VAMP.2 can map using ACME/ARCS (Holyoak and Taghard, 1997), a constraint satisfaction system network (cf. also chapter 2, section 3.2, this book).

An interesting layered computational model of perception and speech recognition and understanding has been proposed: "Models intended to be comprehensive often suppose three or more major layers, often with sublayers, and sometimes with parallel channels that separate and combine to support higher level hypotheses. For example, shading discontinuities and color contrasts may separately support hypotheses about object boundary" (Josephson, 1994b, p. 239)[2]. As a consequence, perception and language understanding can be seen as a form of abduction or at least of "explanatory" reasoning (Charniak and McDermott, 1985; Josephson, 1989a and b, Hobbs, Stickel, Appelt, and Martin, 1993).

Following this theory both perception and interpretation of spoken language must be considered like a process happening in an orderly progression of discrete layers where at each layer an abductive *composite* hypothesis is generated to account for the data presented by the layer below. Separate channels for information (as has been discovered in the case of vision in primates, about shading, texture, and color) provide information to different "abductive" layers. In the case of speech recognition the abductive process that "explains" the acoustic information exploits many kinds of knowledge in the same time: knowledge about the speaker, about the language, about the environment, but also what arrives from other sensorial channel (for instance visual, Massaro, 1987). Abduction is seen as operating on layered information (for instance acoustic, phonetic, phonological-prosodic, syntactical, and pragmatic - Josephson, 1989b). Peirce would have agreed with this point of view, that "recognizing" words in sounds is a kind of inferential process.

The results achieved in the area that refers to the relationships between logic, logic programming, spatial reasoning, and diagrams are also relevant to the problem of visual abductive reasoning (Allwein and Barwise, 1996; Barwise and Etchemendy, 1991; Bennett, 1994; Brown, 1997; Giaquinto,

[2] See also Marr, 1982.

1994; Shanahan, 1995; Shin, 1994). Problems of logic, ontology, and representation in spatial reasoning are also illustrated in Stock (1997).

Finally, interesting studies related to the exploitation of the broader area of visual and spatial reasoning are the ones related to the computational models of diagrammatic reasoning about interaction of objects (Narayanan and Chandrasekaran, 1991; Narayanan and Motoda, 1995), about generalization in geometry (Lindsay, 1994, 1998), and about learning shapes by artificial neural networks (Jacobs and Kosslyn, 1994). From the point of view of cognitive psychology, the empirical results about the relationships between language and spatial models, memory, observation, and the role of maps, environments, and graphs in reasoning are very important (Tversky, 1995, 1997), as well as the role of imagery in reasoning about spatially indeterminate descriptions (Ioerger, 1994)[3].

1.3 Imagery and problem-solving

In accordance with the cognitive-computational architecture previously illustrated, we can consider spatial representations as descriptive. Thus, they are expressed by propositions containing predicates such as spatial relationships and arguments as imaginable objects.

Is there a difference between descriptive and depictive representations?

The spatial representation does not add information that cannot be expressed by propositions; notwithstanding this, the spatial representation is not computationally equivalent to a descriptive one. In several imagery-related tasks (e.g. inspecting) spatial representation may reduce the computational complexity of the solution: the symbolic array adds more constraints to the search. The advantage in processing syllogisms by inspecting a symbolic spatial array is well-known.

Certainly, the symbolic arrays do not have the expressiveness of first-order logic and consequently they cannot represent many situations of vagueness or indeterminacy. Nevertheless they "explicitly" depict the present parts of an image and where they are located relative to one another, as well as depicting what is not present: the full representation is always complete.

As the spatial representations are depictive, and denote the important spatial relations among parts of the image, they are useful in the development of problem-solving devices related to the inspection and transformation of images. Moreover, in processing visual information, two kinds of parallelism are involved: spatial, corresponding to the same operations applied

[3] The role of spatial abilities in the embodied and mediated cognition related to manipulative abduction is illustrated in chapter 3; details on imagination in situated cognition and robotics (cognition as "imagined interaction") are given in Stein (1995).

concurrently to different spatial locations in an image; and functional, which occurs when different operations are applied to the same location.

Cognitive maps can be mentally inspected and used by an agent for route planning; imagery is used in planning and design; as stated above, the use of imagery in scientific discovery illustrates a mechanism of thinking relevant to hypothesis generation; the role of imagery in problem-solving that involves scanning and inspection at multiple levels of structural hierarchy is shown in the computational model by the use of nested arrays of varying granularity, where attention can be focused to retrieve details; finding spatial similarities and equivalence is involved in many problem-solving activities and is related to the problem of visual analogical mapping (Thagard and Hardy, 1992; Thagard, Gochfeld, and Hardy, 1992); imagery also involves the simulation of image transformations in order to anticipate the consequences of an action or event; constructing novel images through operations such as compose, superimpose, and put, allows us to detect information not previously seen.

The main objective is now to consider how the use of visual mental imagery in thinking may be relevant to hypothesis generation. I plan to explore whether the cognitive-computational tool illustrated above can be modified to delineate the first features of what I call *visual abduction*. The following section proposes a cognitive architecture of image-based hypothesis generation in everyday reasoning.

1.4 Visual abduction

1.4.1 Image-based explanation

Having illustrated many issues concerning the phenomenon of imagery, which is commonly and consciously experienced as the ability to form, transform and inspect an image-like representation of a scene, and having indicated that such representations play a role in problem-solving strategies involving visual or spatial properties of an image, I will now discuss, from a computational philosophy perspective, a visual abductive problem-solving strategy. To provide manageable bounds to our very general objective, i.e. to analyze the role of visual hypothesis generation, which is so crucial to scientific discovery, I have initially limited myself to the subtask of illustrating some structurally similar examples from the field of common sense reasoning, where it is very easy to find many cases dealing with what I have just called visual abductive problem-solving. Moreover, I have limited myself to the spatial representation. The spatial representation does not add information that cannot be expressed by propositions; notwithstanding this, the spa-

tial representation is not computationally equivalent to a descriptive one. In several imagery-related tasks (e.g. inspecting) spatial representation may reduce the computational complexity of the solution.

Although there is considerable agreement concerning the existence of a high-level visual and spatial medium of thought as a mechanism relevant to abductive (selective and creative) hypothesis generation, the underlying cognitive processes involved are still not well understood. Notwithstanding this, I will attempt to work around this gap in our understanding: although describing a model able to "imitate" the real ways the human brain works when it makes visual abductions would be best, my primary concern is its expressiveness and inferential adequacy, rather than its explanatory and predictive power as regards psychological research.

Let us consider the following preliminary cognitive case: many visual stimuli are ambiguous, yet people are adept at imposing order on them. As stated in chapter 2 (section 3.2), this is the case when we readily form hypotheses such as that an obscurely seen face belongs to a friend of ours, because we can *explain* what has been observed[4].

More generally, we can face an *initial* (eventually)[5] observed image in which we recognize a problem to solve. For example, given a visual or imagery datum, we may have: 1) to explain the absence of an object; 2) to explain why an object is in a particular position; 3) to explain how an object can achieve a given task moving itself and/or interacting with the remaining objects in the scene/image; 4) to show how we can recognize an object as of significance (for instance the recognition of a stone as a tool, Shelley, 1996).

How can "visual" reasoning perform this explanations? To answer this question it is necessary to show how visual abduction may be relevant to hypothesis generation, that is, how an *image-based explanation* is able to solve the problem given in the initial image.

Faced with the initial image, in which we have previously recognized a problem to solve, as stated above, we have to work out an *imagery hypothesis* that can explain the problem-data[6]. Thus, the formed image acquires a hypothetical status in the inferential abductive process at hand.

1) We have to *select* from long-term memory a visual (imagery) description that is able to explain the anomaly that needs to be solved; 2) we have to justify the presence/absence of a given object in a scene selecting a suitable imagery explanatory hypothesis; 3) for instance we have to visually solve the

[4] We have also illustrated that Peirce's ideas on perception are related to this example.

[5] Of course the initial spatial image can be the representation of a real case.

[6] When discussing some problems related to the abductive reasoning, Bayesian networks, perception, and vision, also Poole (2000) underlines that in vision we can think of a scene *causing* the image: "the scene produces the image, but the problem of vision is, given an image, to determine what is in the scene", that it is an abductive task.

well-known monkey-banana problem (cf. the following section): every formed visual representation of the effect of a sequence of actions the monkey can perform may be considered as a hypothesis generation. Such an hypothesis, if successful, is viewed as the one selected that gives a solution of the problem; 4) a slightly different selected version of the initial image can perform the task of giving sense to an object.

The generation of a "new" imagery hypothesis can be considered the result of the *creative* abductive inference previously described; in this respect we can consider how the imagery representations of new hypotheses lead to scientific discovery. The selection of an imagery hypothesis from a set of pre-enumerated imagery hypotheses, stored in long-term memory, also involves abductive steps, but its creativity is much weaker: this type of visual abduction can be called *selective* (see chapter 2, section 1.2).

All we can expect of visual abduction is that it tends to produce imagery hypotheses that have some chances of turning out to be the best explanation. Visual abduction will always produce hypotheses that give at least a partial explanation, and therefore have a small amount of initial plausibility. In this respect abduction is more effective than the blind generation of hypotheses. How can we perform the generation of imagery hypotheses which are able to explain problem-data? This complex task can be achieved in an environment supplied with suitable levels of expressivity and adequate inferential model.

1.4.2 Imagery hypotheses

In the four simple examples stated above we were faced with an initial image in which we had 1) to explain why an object is in a particular position; 2) to explain the absence of a particular object; 3) to explain how an object can achieve a given task moving itself and/or interacting with the remaining objects in the scene/image; 4) to show how we can recognize an object as of significance (for instance the recognition of a stone as a tool, Shelley, 1996).

We shall clarify the first case with the following example, which is based on common sense reasoning: we see a broken horizontal glass on the floor, near the table. On the floor there are also some leaves and we see that the window is open. If we retrieve from long-term memory another visual (imagery) description still containing the glass (intact), the table, and the window, and we recognize this new representation as being a slightly different version of the previous one, we have to explain the presence of the leaves and broken glass in the initial image. They constitute an *anomaly* (see chapter 6, section 2) that needs to be solved (explained). If we are able to link the leaves to the presence, say, of wind, we are in turn transported to a new imagery explanatory hypothesis. Of course such a performance needs a preliminary knowledge base and a suitable form of connections between data.

We can perform these connections by means of a neural network that links objects in the spatial world universe. In particular we may be able to establish excitatory links between objects belonging to the same arrays that are stored in long-term memory: this mechanism will enable the detection of the anomaly represented by the leaves and broken glass in the example above. A further link established between the leaves and wind will lead to the completion of the explanatory task.

The second case deals with the capacity to justify the absence of a given object in a scene. Let us consider the following example: one of our friends is accustomed to travel the same route every day. The road passes near to a little bridge, under which ducks can usually be "seen" swimming. On a particularly cold day our friend does not see the ducks (the detection of this fact visually derives from the comparison of the observed image with a similar mental one - containing the ducks! - pre-stored in long-term memory). He asks himself where the ducks could be, but, since he has never seen any ducks in a different setting, while he is able to detect the anomaly he is unable to explain it. The imagery explanatory reasoning is impossible: therefore, it is stopped.

On the contrary, if our friend had previously seen the ducks, say, under the roof of a farmhouse, once he has visually detected the absence of the ducks he can retrieve from long-term memory the image of the ducks sleeping under the roof. The imagery explanatory hypothesis is immediately achieved. The computational mechanism is the same as in the first example. By comparing an image of a place with a similar one stored in long-term memory it is possible to detect an anomaly (such as the absence of the ducks). A link is established between different objects (in the example, between the ducks and the roof) which allows the retrieval from long-term memory of another image (spatial world) that constitutes an hypothesis that explains the absence of the ducks.

The third case deals with the well-known monkey-banana problem. In a room there is a banana, a box, and a monkey. The monkey cannot reach the banana because it is on the ceiling, but it can push the box to a point below the banana, climb on top of it and so reach the banana. Every visual representation of the effect of a sequence of actions the monkey can perform may be considered as an hypothesis generation. Such an hypothesis, if successful, is viewed as the one that gives a solution of the problem. This previous example satisfies cases in which we have *to explain*, in an initial (eventually) observed image, how an object can achieve a given task, moving itself and interacting with the remaining objects in a scene.

The last case deals with a process of "interpretation" of the meaning of a given object.

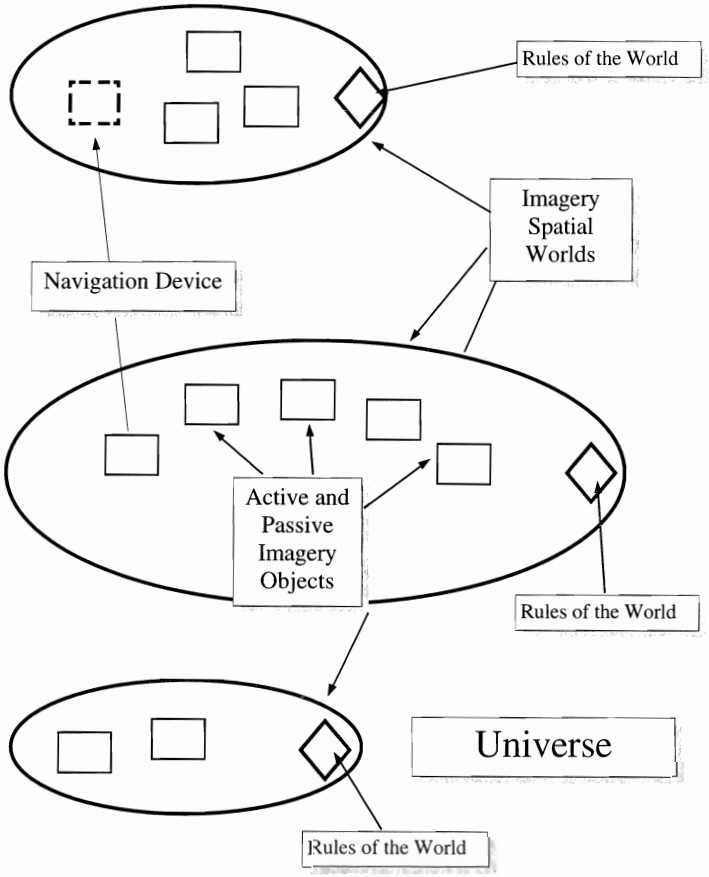

Figure 3. Visual abduction system structure.

1.4.3 Visual abduction system structure

I will propose the cognitive architecture of a system that focuses on spatial reasoning tasks and is hierarchically organized into three levels: 1) a *universe* of spatial worlds 2) *spatial worlds* 3) *imagery objects* that are included as elements in each spatial world (an imagery object may be included in one or many worlds). At the computational level, the universe has to be composed of a collection of representations that describe each spatial world, and of a *navigation device* that allows an imagery object to move from the originary world to a different world. Thus, the different worlds can communicate

with each other by means of the navigation device. Indeed, each world (represented, according to Glasgow and Papadias computational model, by a nested symbolic array) is composed of a domain of suitable imagery objects (represented by the symbols in the array) and by a collection of rules identifying it (e.g. Newtonian mechanics).

The imagery objects are of two kinds: *active* and *passive*. Both kinds of objects "know" the rules of the world where they are included (or to which they move); moreover, they can be individually characterized by a family of functions that imply some definite kinds of action they are able to perform. Finally, 1) each active object may interact with the remaining objects in a world; and 2) each passive object in a world may only act after interacting with an active object (Figure 3). To represent quantitative relations explicitly among the objects of the array we may modify Glasgow and Papadias' knowledge representation scheme by means of topological combinatorial considerations upon the hierarchy of spatial worlds. Moreover, it is necessary to add the representation of empty entries and of suitable distance functions.

Since the universe of spatial worlds can be mathematically represented as a tree, we can consider its adjacency matrix. A well known theorem in graph theory allows us to manage easily paths of any length and therefore to consider a lot of relations beyond the mere inclusion (e.g. "to be next to but not included", "to be elsewhere but not far from here" and so on). Such a capability induces the possibility to match spatial worlds consistently with these new relations, in order to provide a detailed computational problem-solving strategy for solving visual abduction problems described above (that is the explanation of the absence or of the position of an object).

It is possible to develop a set of functions with several capabilities, such as object manipulation, array retrieval, mapping and so on. The principal interest is to design functions of cognitive importance in array theory. The complete set of such functions covers the functionality of the *navigation device*. Here are some examples.

Consider a hierarchy of nested arrays (Figure 4). We are trying to give a precise meaning to "cognitive" sentences like "Array A is more interesting than array C", "There is a stronger link between G and F than between G and I" or "The object represented by array A is better known that any other". Someone could wonder why we think array A is more interesting than C, seeing that we do not know what the arrays A and C are representing. We think that such considerations can be simply suggested by the topological structure of the hierarchy of the arrays (that is, actually, a tree).

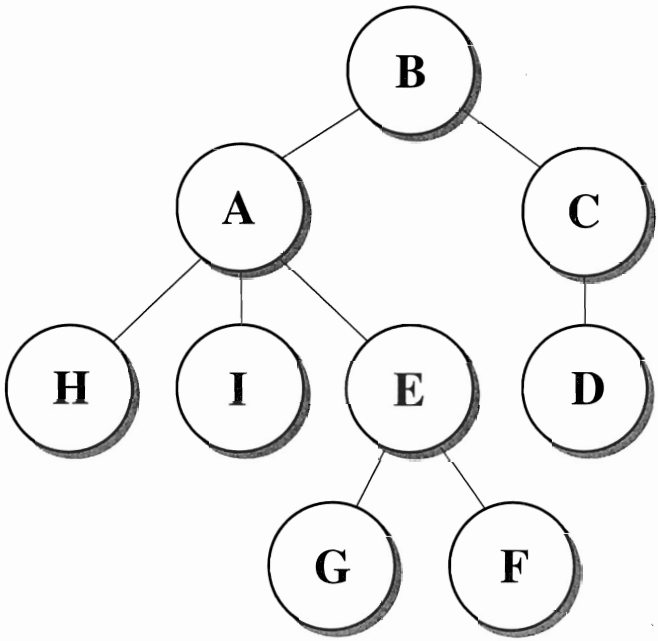

Figure 4. Hierarchy of nested arrays.

The basic idea is to take into account not only paths of length one between the nodes (e.g. the path E-F in the above illustration), but also paths of length more than one (e.g. the two-lenght path G-E-F). Obviously, from a cognitive point of view, a path of length one is more "important" than a path of length two. In order to consider this situation we give different weights to n-length paths (for instance a n-length path might have a weight of $1/n!$ - one over n factorial -). So the path between G and F will be weighted 0.5 while the path between G and I will be weighted roughly 0.166 giving a justification of one of the above sentences. The next step is to answer a (seemingly) meaningless question: "How many paths are there between A and A?". The answer is: "There are four paths of length two, seven paths of length four and so on". We claim that the sum of the weights of such paths could be a measure of the cognitive interest of the array. This allows us to give a meaning to sentences like "Array A is the most important".

The cognitive meaning of a two length path (namely A-I-A) could be represented with the following sequence of operations: retrieval of the image A - focusing of a particular (namely I) - return to the image A. The topological structure of the representation is clearly subjective, reflecting one's points of

view. For instance I know my home much better than my neighbour's one, correspondingly the array representation of my home is far richer.

To solve the third visual abduction problem introduced above (monkey-banana problem) it is easy to use a classical means-ends analysis strategy. The room is represented by an array where all the objects are indicated by symbols on the array. If we describe a primitive distance function, say dist(x,y), we are immediately transported to a new formulation of the goal, that is, dist(monkey,banana)=0. Let us suppose the "object" monkey is supplied with a set of primitive functions, such as *walk-right*, *walk-left*, *push-left*, *push-right*, *climb*, *grasp*. It is possible to describe visually the effect of every function in the array. A feasible procedure is 1) choose an action 2) if this action leads to a decrease in the distance then 2.1) do it 3) else go to 1). Such an algorithm clearly does not always lead to the goal. It has been improved with a device that is capable of detecting stationary states which are not the solution of the problem. We can also avoid this impasse by adding to the distance function a penalty function which evaluates the "distance" of the actual state from the set of stationary states.

If we consider each new step leading to the goal (i.e. each new configuration of the array generated by the new positions of the objects) as an *imagery state*, we can say that the "monkey" (or "people" faced with the monkey-banana problem) forms different imagery hypotheses devoted to achieving the task. Thus, each step represents a particular imagery world. The configuration of the array in which the goal is achieved performs the image-based best explanation, i.e. the visual abduction: this generated abductive imagery hypothesis is the best explanation of the problem-data, and hence able to perform the planning task (Figure 3).

Finally, if our spatial world represents a room, it can be supplied with a collection of rules of varying detail (e.g. Newtonian mechanics), yet it can also be supplied with completely different rules. Of course, using less rules results in the objects of the world being less constrained. For example, weakening the rules of Newtonian mechanics can lead to a new kind of spatial world, which is very abstract and considered as being a "virtual" one (Degli Antoni and Pizzi, 1991; Wilson, 1997).

From a functional point of view the proposed architecture is extendable in several ways - each related to different capabilities. First, we may limit ourselves to solving problems in a unique spatial world; this approach naturally leads to qualitative physics reasoning. In the examples above the initial and final states were slightly different: we can also consider strategies that involve a greater number of different spatial worlds. Nevertheless, it is possible to build new imagery worlds by moving objects in the universe using the navigation device and then mapping the generated worlds: my architec-

ture can be extended in order to map worlds onto worlds, that is, imagery hypotheses onto imagery hypotheses.

How are visual and temporal abduction (cf. the following section) related to each other? First of all we can study a particular sequence of spatial worlds as forming a kind of *history* that leads to an observed (or imagery) *shape* or to an observed (or imagery) *structure*. In this last case we deal with the visual abduction that involves primarily the role of shapes and structures. I think that other interesting results could come from research concerning the possibilities of expressing spatiality using temporal reasoning and vice versa.

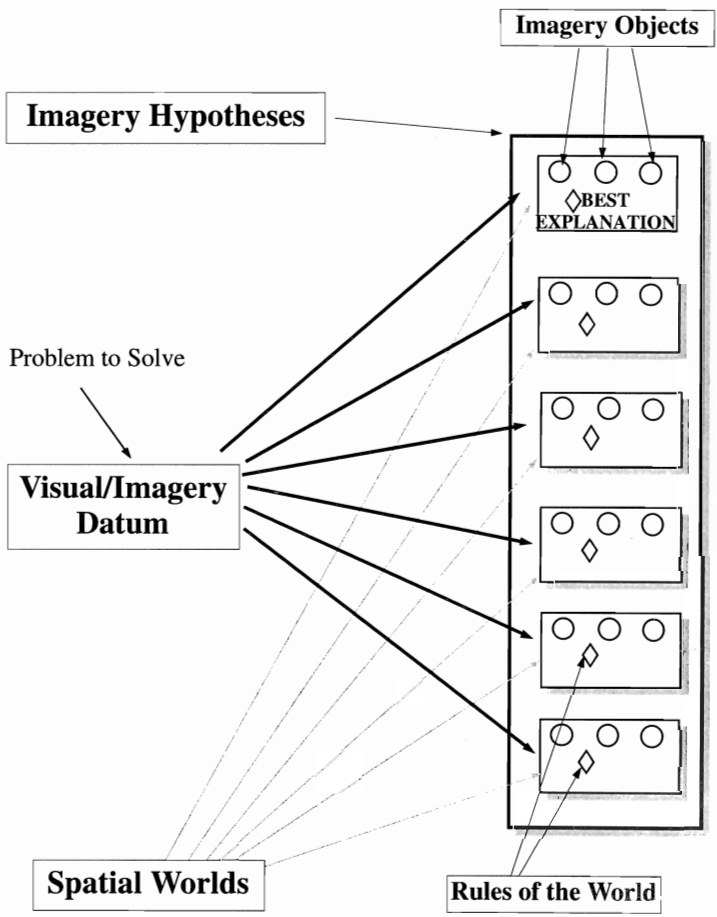

Figure 5. Visual abduction as image-based explanation.

I plan to enrich the architecture to represent cases that deal with more dynamic situations such as forming new images 1) changing in dimensions, 2) changing in shape; 3) changing in structure (discrete vs. dense; bounded-unbounded; number of objects). We find these last kinds of image trasformations in many cases related with *creative* visual abduction, as involved in scientific discovery, where we have also to establish relevant strategies for passing from the *propositional-verbal* level to the *visual* one and vice versa[7].

This section has examined a knowledge representation scheme together with certain cognitive results in order to explore whether a kind of hybrid imagery-linguistic representation can be improved and used to model image-based hypothesis generation.

2. TEMPORAL ABDUCTION

The aim of this section is to highlight the role of time in abductive reasoning. The notion of time is present in any intelligent activity: the concept of time is profoundly involved in human perception and understanding of the world. Things appear to remain in a particular state for a period of time until a certain event happens, so we can say that time is central to reasoning about change and action.

Temporal reference is highly integrated in scientific knowledge but this is also the case in human common sense reasoning, both verbal and at the level of visual imagery. Humans are involved in managing time when coordinating with the environment; memories and many mental models seem to be organized around time - past events come to mind when reconstructed with the help of a time framework (chronologically). After illustrating some philosophical and epistemological aspects that concern the fundamental relationship between science and time - Prigogine's "forgotten dimension" - I will explore the role of time representations in the field of computational philosophy, to stress some initial features of abduction and nonmonotonicity in temporal reasoning.

2.1 Temporal reference

The notion of time is present in any activity that involves intelligence: the concept of time is deeply involved in human perception and understanding of the real world. Things appear to remain in a particular state for a period of time until a certain event happens, so we can say that time is central to rea-

[7] For example, Johnson-Laird (1983) defines a temporal model as a sequence of spatial models. Hence, temporal reasoning is also involved in the field of so-called visual thinking, though this is outside the scope of this chapter.

soning about change and action: the first task is to consider the different states or conditions of a thing and define how they are related. Moreover, we can say that we have knowledge about a causal relationship when we can use it to define how the related system evolves.

It is well known that temporal reference is highly integrated in scientific knowledge but this is also the case in human common sense reasoning, both propositional and imaginary. Humans are always involved in managing time when coordinating with the environment in everyday life; memories seem to be organized around time - many past events come to mind when reconstructed with the help of a temporal framework (chronologically). Our idea of "time-flowing-forward" (along a single line) is too strong, and contradicts counterintuitive concepts of time that deviate from it such as "some interpretations of quantum mechanics which perceive time as constantly splitting into the future" (Shoham, 1988, p. xiii), time in relativity theory or some philosophical speculations about circular time (see section 2.3).

Many studies in the various areas within artificial intelligence involve reasoning about time:

1. medical diagnosis: the systems have to determine when illness symptoms occurred as well as their evolution, in order to detect correctly the cause and possible suitable treatment;
2. a planning system has to consider the duration of the actions and the task to be achieved, and has to manage these relationships over time;
3. representing agents knowledge has to deal with beliefs changing as a result of external input and internal reasoning;
4. to control a process in industrial process supervision requires the consideration of past states, historic evolution and when certain operations will affect the evolution of the process;
5. natural language understanding has to consider the treatment of the verb tense, which is a fundamental element in the meaning of sentences.

Thus, the following section will illustrate some epistemological aspects that concern the fundamental relationship between science and time; section 2.3 will explore the role of time representations delineated in the field of computational philosophy, to highlight some initial features of what I call *temporal abduction.*

2.2 Science and time: the forgotten dimension

Prigogine introduces the general problem of time in science as follows (Prigogine, 1980, p. xi):

> This book is about time: I would like to have named it *Time, the Forgotten Dimension*, although such a title might surprise some readers. Is not time incorporated from the start in dynamics, in the study of motion? Is not time the very point of concern of the special theory of relativity? This is certainly true. However, in the dynamical description, be it classical or quantum, time enters only in a quite restricted way, in the sense that these equations are invariant with respect to time inversion, $t \rightarrow - t$. Although a specific type of interaction, the so-called superweak interaction, seems to violate this time symmetry, the violation plays no role in the problems that are the subject of this book.

It is well known that Newton's concept of time did not challenge the Aristotelian concept. Newton assumed that concepts of space and time were "well-known to all", but he rejected the common people's conception of space and time in terms of relations to sensible objects. Time is considered as universal and absolute, as a *variation* in the state of motion; as a parameter it is unaffected by the transformations that it describes. He asserted: "Absolute time, and mathematical time, of itself, and from its own nature, flows equally without relation to anything external. [...] Absolute space, in its own nature, without relation to anything external, remains always similar and immovable" (Newton, 1934, p. 6).

It is well known that Einstein abandoned the standard concept of *simultaneity* of events happening at the same time in an absolute sense to replace the old received views. Minkowski developed the special theory interpreting it using a four-dimensional space-time continuum, providing a significant contribution to the development of the general theory. Einstein came to think of time as like the three spatial dimensions, so space and time were no longer considered as separate from each other. Newton thought that temporal concepts such as duration and simultaneity were unproblematic, but Einstein argued that whether two events are perceived as simultaneous depends on the reference frame of the observer. So we have to think of time as a part of space-time, because it has no meaning independent of space-time.

The renunciation of absolute time had a relevant impact on philosophy, which can be seen by studying the writings of Bergson, Croce, Lovejoy, Ortega y Gasset and Whitehead. The idea of the direction of time (*durée*, *élan vital*) is present in Bergson's philosophy; unfortunately, the attempts to answer unscientifically the epistemological question concerning the one-sidedness of time by naive generalizations and analogies which emphasize "feeling" rather that cognition led to the well-known forms of metaphysical evolutionism (see Le Roy's critique of science in order to argue for a spiritualist "philosophy of freedom and for the superiority of intuition and religious experiences over science" - LeRoy, 1899).

Hence, philosophies of time and freedom can be considered as many ways of replying to the determinism generated by physics; in these philosophies change, irreversible, poietic, and time-directed duration, pertain to the field of irrationality. Even in the postpositivistic philosophy of science the epistemological problem of scientific creative thought, of discovering new ideas (Popper, Kuhn, Feyerabend), is confined to the realm of a strange, eccentric, and irreversible temporality of "intuition": the actual genesis of discovery eludes a real scientific study for the reason that it is outside the framework of logic. Popper considers that "there is no such thing as logical method of having new ideas or of a logical reconstruction of this process and that every discovery contains "an irrational element" or "a creative intuition" in Bergson's sense of these terms (Popper, 1959, pp. 31-32).

However, the concept of evolution emerged with force in the nineteenth century in physics (related to the concept of irreversibility), biology, and sociology (related to the concept of history). In physics it was offered by the second law of thermodynamics, the law of increase of entropy. Notwithstanding this, evolution and irreversibility continued to appear as an illusion related to the complexity of the combined behavior of simple microscopic objects[8]. As for the present, we know that these objects (for instance, elementary particles) can be produced and can decay. Moreover, irreversible processes may play a poietic role in nature, which appears very clearly on a biological level.

Scientists always knew that a description in which past and future play the same role does not apply to all phenomena. Notwithstanding this, the phenomena characterized by the *direction of time* were excluded from classical, relativistic, and quantum physics. All time-oriented processes (for instance, biological) were regarded as an effect of special, improbable initial conditions, corresponding to supplementary approximations that we superpose upon time-reversible laws. As stated by Prigogine:

> [...] living organisms are far-from-equilibrium objects separated by instabilities from the world of equilibrium and [...] living organisms are necessarily large, macroscopic objects requiring a coherent state of matter in order to produce the complex biomolecules that make the perpetuation of life possible (1980, p. xv).

Prigogine demonstrates many simple irreversible processes, such as heat conduction, but also complicated processes involving *self-organization*, and the central role of *fluctuations*: far from equilibrium, chemical systems that include catalytic mechanisms may lead to dissipative structures and, once a

[8] As Einstein stressed in a letter to his friend Besso: "There is no irreversibility in the basic laws of physics. You have to accept the idea that subjective time with its emphasis on the now has no objective meaning" (Einstein and Besso, 1972, quoted in Prigogine, 1980).

dissipative structure is formed, the homogeneity of time and space may be destroyed. Bifurcation phenomena too, that imply history in physics and chemistry and, consequently, some preferential directions of time, may be mathematically described (Thom, 1975).

2.3 Computational philosophy of time

2.3.1 Reasoning tasks

A theory of time has to deal with i) a *formal language* for describing what is true and what is false over time, what changes and what remains constant, and ii) a *set of rules* governing change. In temporal reasoning we can define four general tasks (Shoham, 1988):

1. Given a description of the world over some period of time, and a set of rules about change, to predict the state of the world in a given future time.
2. Explanation: to generate a description of the world at some past time which accounts for the world being the way it currently is.
3. Learning about physics: given a description of the world at a given time, to generate a set of rules that can govern change and that account for the regularities in the world.
4. Planning: given a description of some desired state of the world over a period of time, and a set of rules, to generate a sequence of actions that would result in a world fitting that description.

2.3.2 Logic and ontology of time

Recently, temporal logic has been applied to computer science. To formalize the concept of time we need first a language that is able to describe what is true and what is false over time and, second, that is able to describe expressions like "time A is before time B". Among the entities of the language we have to distinguish between events, facts, and processes. For instance, facts and properties are things that are true over time: "Ann is tall". Events are things that happen, for instance "to start playing a song".

Temporal logic has its origin in philosophy and was studied to investigate the general structure of time (Newton-Smith, 1980; Prior, 1955; van Benthem, 1983). McDermott's temporal logic (1982) and Allen's theory of action (1984) are more oriented toward formalizing time in common sense reasoning. In logic there are three ways to manage time: first order logic with temporal arguments (where functions and predicates are extended with the

additional temporal argument, the constant t_0 - present time - and a temporal ordering $\leq$,) modal temporal logics (with temporal operators and where each possible world represents a different time), reified logics (where we work in a meta-language in which a formula in the object language - the classical first order logic - becomes a term and where we are able to talk about things that are true over time).

At the ontological level we may have different primitives like time points (instants) and/or time intervals (periods) that may have different properties (discrete vs. dense, bounded vs. unbounded, precedence: linear, parallel, branching, circular). For instance we can define time by a dense set of instants over which an ordering relation $\leq$ is introduced which is reflexive, anti-symmetric, and transitive. Allen defined an *Interval Calculus* involving 14 relations corresponding to mutual relationships between two periods (Allen, 1984).

2.3.3 Prediction and the qualification problem

When we have to predict the collision of two rolling balls, using classical mechanics, given for instance their current trajectories, we do not state that there are other balls (or winds, or that the balls are about to explode, and so on) that influence the trajectory of the two particular balls. This kind of description is not sufficient when submitted to formal inference techniques. A classical example is done in the area of common sense reasoning by the Yale shooting problem (Hanks and McDermott, 1987). When one pulls the trigger of a loaded gun one would like to predict that a loud noise will follow, yet there are many conditions that have to be verified - that the gun has a firing pin, that there is air to carry the sound, and so on. This is the well-known *qualification problem* (Shoham, 1985, 1988, 1989, Shoham and McDermott, 1988): the problem of making predictions about the future without considering everything about the past. As clearly stated by Shoham, the most serious disadvantages of the classical mechanics approach are:

> 1) The initial conditions must refer to a unique instant time. Furthermore, we must give a complete description of the initial conditions, which, unless abbreviated, is too costly. So far we have no way of saying "this is all the information that is relevant to the problem". 2) The physics specify which predictions are warranted by the initial conditions, but not how to make them. This information must be supplied from outside the physics. 3) The rules of physics are constraints on the simultaneous values of quantities. This instantaneous flavor of the rules, which makes the formulation elegant and parsimonious, is the reason that predictions about extended periods of time are hard to make (Shoam, 1988, p. 9).

Hence, classical mechanics is a bad model for temporal reasoning[9]. The overidealization of the initial conditions immediately leads to the qualification problem, consequently we have to be aware that we may make predictions based on very partial information, hoping that the ignored factors will not get in the way (*chronological ignorance*). So we have to be prepared to make mistakes in our predictions and to withdraw them.

Moreover, although we may be able to make predictions about short future intervals, in a realistic complex environment we have to deal with the length of time intervals in the future to which the predictions refer. We are sure that the best predictions deal with the shortest interval of time (if possible, instantaneous). This is the *extended prediction problem* (Shoham, 1988, 1989). Finally, when we predict that a fact will remain unchanged further difficulties arise. This is the so-called *persistence problem* (McCarthy and Hayes, 1969). Shoham is persuaded that the persistence problem and the well-known frame problem coincide:

> For example, if the action PAINT(HOUSE17,RED) is taken in any situation s_1, the result is a situation s_2 in which the color of HOUSE17 is red. But now consider taking the action REARRANGE-FURNITURE in s_2, which results in a new situation s_3. What is the color of HOUSE17 in s_3? One would like to say that it is still red, since rearranging the furniture does not affect the color of the house, but unfortunately the formalism does not say that. We could add to the formalism the fact that after you rearrange the furniture the color remains unchanged, and this would be what McCarthy and Hayes call *frame axiom*. The problem is that you'd need many such axioms: rearranging the furniture doesn't clean the floors, doesn't change the President of the United States, and the list can be continued infinitely [...] McCarthy and Hayes called this the *frame problem* (Shoham, 1989).

In the qualification problem and in the extended prediction problem we have to ignore much of the information that is potentially relevant, and thus we have to retract some of the conclusions in the face of new evidence. This kind of defeasible inference was studied in the field of *nonmonotonic logics*[10]. In these logics we may "jump to the conclusion" in the absence of evidence to the contrary: inferring that a thing can fly from the fact it is a

[9] Other considerations concerning temporal reasoning are given in the area of the so-called dynamical approach to cognitive systems, which stresses that all cognitive processes have to be studied taking into account the "real" time within they happen (Port and van Gelder, 1995) (cf. also chapter 3, section 4, this book). An interesting illustration of the research about the relationships between naive time, scientific time, temporal patterns, and recognition of auditory patterns in time, is given in Port, Cummins, and McAuley, 1995.

[10] Classical contributions concerning nonmonotonic reasoning can be found in Ginsberg (1987) (cf. also chapter 2, section 1.2, this book).

bird, but withdrawing that inference when an additional fact is added, i.e. that the thing is a penguin.

Hence, temporal reasoning frequently involves the problem of retracting the conclusion. Lets consider the Yale shooting problem illustrated above: it concerns a scenario where Fred is alive, a gun is loaded and shooting it would kill Fred (Vila, 1994). A universal frame axiom states that anything that is not affected by an action, persists through the execution of the action. From this initial situation we have first a waiting action and then the gun is fired. Is Fred still alive?

The Yale shooting problem can be formalized as follow:

HOLDS(alive(Fred),loaded(gun); S1)
∀S.HOLDS(loaded(gun); S)→
HOLDS(¬alive(Fred);Result(shooting, S))
S2 = Result(wait, S1)
S3 = Result(shooting, S2)
HOLDS(alive(Fred); S3)? (Vila, 1994)

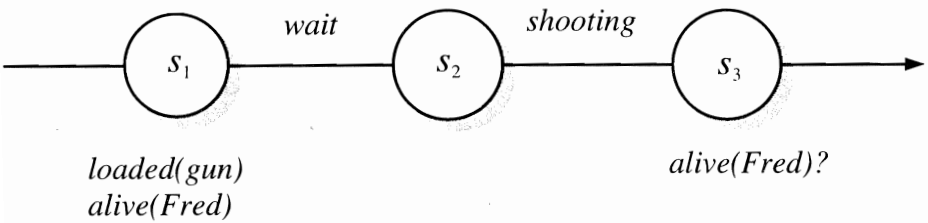

Figure 6. Yale shooting problem.

The obvious answer is that Fred is dead. Nevertheless, the classical systems are only able to conclude that either Fred is dead or the gun has been unloaded by the waiting action (Figure 6).

Several authors developed formalisms to overcome the frame problem embedded in this scenario: pointwise circumscription (Lifschitz, 1986), logic of persistence (Kautz, 1986), chronological minimization (Shoham, 1988). Evans used negation as failure (Evans, 1989). For instance, minimization consists, intuitively, of delaying as much as possible the occurrence of the abnormality.

Using a version of event calculus, Shanahan (1990) presents a solution of the so-called bloodless variation of the Yale shooting problem that uses abduction and employs circumscription to achieve default persistence. The ontology includes events, time points, and properties. In the bloodless variation of the Yale shooting problem the history of events - load then wait then

shoot - is extended by "alive", that holds afterwards. This history can be formalized as follows:

1. $happens(e_0)$
2. $time(e_0,t_0)$
3. $act(e_0,birth)$
4. $happens(e_1)$
5. $time(e_1,t_1)$
6. $act(e_1,load)$
7. $t_1 > t_0$
8. $happens(e_2)$
9. $time(e_2,t_2)$
10. $act(e_2,wait)$
11. $t_2 > t_1$
12. $happens(e_3)$
13. $time(e_3,t_3)$
14. $act(e_3,shoot)$
15. $t_3 > t_2$
16. $t_4 > t_3$ (Shanahan, 1990).

First, we may think that the presence of the new fact

$[holds\text{-}at(alive, t_4)]$

simply generates a different prediction and compels the withdrawal of the previous one, i.e. that *dead* holds as a result of the shoot event. Second, we may also consider the new fact as requiring an *explanation*, and then we have only to derive new predictions from the explanation. In this second case, since the shoot event happened before t_4, and since shooting terminates *alive*, either the shooting was unsuccessful (i.e. either the wait event or some other event not indicated in the history unloaded the gun), or the new fact can be explained by a new kind of event that initiated alive, for instance a resurrection event.

Of course to explain the new fact it is possible to *abduce* many tentative hypotheses, that generate many *amended histories*, that in turn make possible new predictions. Hence, we are faced with a cycle composed of an abduction for the explanation and a deduction for the prediction. Shanahan solves the whole problem by providing a formalism that is able to choose the preferred explanation by making domain predicates abducible or not. In the

first case the preferred explanation is that the wait event unloads the gun. In the second case the preferred explanation for

$[holds\text{-}at(alive, t_4)]$

involves reincarnation:

$\{happens(e), time(e,t), act(e,birth), t < t_4, t_3 < t\}$.

Further complications arise where there are many preferred explanations for a fact.

The results of these studies stress that assumptions - the essence of nonmonotonic reasoning - can also be considered on temporal extensions. Changes in the truth value of a proposition or in its temporal extensions may involve the temporal extensions of the truth of a related proposition. For instance, if I know that "the band played my favorite song" I can predict approximately three minutes. But, if suddenly, a terrific earthquake happens, then the estimated playing period is reduced.

This section has examined the concept of time in science together with certain results in order to show the complexity of the field of temporal reasoning.

First, I described the rejection of the classical mechanical concept of absolute time in favour of Einstein's special theory, and also the rejection of reversible time, due to the investigation of dissipative structures (objects far from equilibrium, separated by *instabilities* from the world of equilibrium, involving the theme of time oriented, self-organized, processes), and to the general domain of evolving processes (historical, biological, cosmological). Second, faced with the general problem of managing time in human reasoning, we have also found that classical mechanics is a bad model: the initial conditions must refer to a unique time instant and we must give a complete description of the initial conditions. Moreover, also in common sense temporal reasoning the information is characterized by instabilities that refer to its indeterminacy, incompleteness, and mutability. To provide manageable bounds to the objective of studying temporal reasoning it is necessary to exploit the resources of temporal logic and nonmonotonic logical models. Abduction is also important when, in a particular history, we must deal with the problem of generating an explanation of an unexpected fact and then making a prediction of the possible evolution[11].

[11] Other approaches to the problem of the logical, computational and cognitive representation of time are given in the Proceedings of the AAAI-94 Workshop "Spatial and Temporal Reasoning", together with some suggestions about the possibilities of representing relationships between space and time in cognitive processes.

Chapter 6

Governing Inconsistencies

1. ROADS TO CHANGES IN THEORETICAL SYSTEMS

We have seen that for Peirce abduction is an *inferential process* in a very particular sense (cf. chapter 1, sections 1.2 and 3.2), abduction

> is logical inference [...] having a perfectly definite logical form. [...] The form of inference, therefore, is this:
> The surprising fact, *C*, is observed;
> But if *A* were true, *C* would be a matter of course,
> Hence, there is reason to suspect that *A* is true (*CP* 5.188-189, 7.202).

C is true of the actual world and it is surprising, a kind of state of doubt we are not able to account for using our available knowledge. Philosophers of science in this century have illustrated that inconsistencies and anomalies often play this role of surprise in the growth of scientific knowledge.

Hence, contradictions and inconsistencies are fundamental in abductive reasoning, and abductive reasoning is appropriate for "governing" inconsistencies: this chapter illustrates *abductive reasoning* and its formal models in order to classify and analyze the different roles played by *inconsistencies* in different reasoning tasks and scientific discovery. The aim is to identify aspects of inconsistencies not covered by certain formalisms and to suggest extensions to present thinking, but also to delineate the first features of a broader constructive framework able to include abduction and to provide constructive solutions to some of the limitations of its formal models. There

are many ways of "governing" inconsistencies (cf. chapter 2, section 2): from the methods activated in diagnostic settings and consistency-based models to the typical ones embedded in some forms of creative reasoning, from the interpretations in terms of conflicts and competitions to the actions performed on empirical and conceptual anomalies, from the question of generating inconsistencies by radical innovation to the connectionist treatment of coherence. The conclusions presented here aim to represent a step forward in the understanding of the use of inconsistencies in scientific creativity.

In different theoretical changes we witness different kinds of discovery processes operating. Discovery methods are *data-driven* (generalizations from observation and from experiments), *explanation-driven* (abductive), and *coherence-driven* (formed to overwhelm contradictions) (Thagard, 1992). Sometimes there is a mixture of such methods: for example, a hypothesis devoted to overcome a contradiction is found by abduction. The detection of an anomaly demonstrates that an explanation is needed. The next move of the process of explanation is to obtain a possible explanation. Therefore, contradiction and its reconciliation play an important role in philosophy, in scientific theories and in all kinds of problem-solving. It is the driving force underlying change (thesis, antithesis and synthesis) in the Hegelian dialectic and the main tool for advancing knowledge (conjectures and refutations - Popper, 1963 - and proofs and counter-examples - Lakatos, 1976 - in the Popperian philosophy of science and mathematics)[1].

Following Quine's line of argument against the distinction between necessary and contingent truths (Quine, 1979), when a contradiction arises, consistency can be restored by rejecting or modifying any assumption which contributes to the derivation of contradiction: no hypothesis is immune from possible alteration. Of course there are epistemological and pragmatic limitations: some hypotheses contribute to the derivation of useful consequences more often than others, and some participate more often in the derivation of contradictions than others. For example, it might be useful to abandon, among the hypotheses which lead to contradiction, the one which contributes least to the derivation of useful consequences; if contradictions continue to exist and the assessed utility of the hypotheses changes, it may be necessary to backtrack, reinstate a previously abandoned hypothesis and abandon an alternative instead.

Hence, the derivation of inconsistency contributes to the search for alternative, and possibly new, hypotheses: for each assumption which contributes

[1] Also psychoanalysis relates creative thinking to something contradictory: creative expression is explained in terms of sublimation of unconscious *conflicts*, as Freud demonstrated in his famous analysis of the symbolic meanings of the works of Leonardo da Vinci (Freud, 1916).

to the derivation of a contradiction there exists at least one alternative new system obtained by abandoning or modifying the assumption.

Anomalies result not only from direct conflicts between inputs and system knowledge but also from conflicts between their ramifications: "noticing a particular anomaly may require building long inference chains tracing ramifications until a contradiction is found" (Leake, 1992, p. xiii). Any explanation must be suitably plausible and able to dominate the situation in terms of reasonable hypotheses. Moreover, the explanation has to be relevant to the anomaly, and resolve the underlying conflict. Finally, in some cases of everyday (and practical) anomaly-driven reasoning the explanation has to be useful, so it needs information that will point to the specific faults that need repair.

The classical example of a theoretical system that is opposed by a contradiction is the case in which the report of an empirical observation or experiment contradicts a scientific theory. Whether it is more beneficial to reject the report or the statement of the theory depends on the whole effect on the theoretical system. It is also possible that many alternatives might lead to non-comparable, equally viable, but mutually incompatible, systems[2].

Why were the photographic plates in Röntgen laboratory continually blackened? Why does the perihelion of the Mercury planet advance? Why is the height of the barometer lower at the high altitudes than at the low ones. These are examples of problems that come from observation, but they are problematic in light of some theory, that is unexpected and anomalous. The first was problematic because it was tacitly supposed at that time that no radiation or emanation existed able to penetrate the container of the photographic plates; the second because it conflicted with the Newtonian theory; the third was problematic for the supporters of Galileo's theories because it contradicted the belief in the "force of vacuum" that was adopted as an explanation of why the mercury does not fall from a barometer tube (Chalmers, 1999).

Dealing with the problem of withdrawing scientific *paradigms* Kuhn writes:

> Discovery commences with the awareness of anomaly: i.e., with the recognition that nature has somehow violated the paradigm-induced expectations that govern normal science. It then continues with a more or less extended exploration of the area of anomaly. And it closes only when the paradigm theory has been adjusted so that the anomalous has become the

[2] Thagard proposes a very interesting computational account of scientific controversies in terms of so-called *explanatory coherence* (Thagard, 1992) (cf. also chapter 2, section 1.3, this book), which improves on Lakatos' classic one (Lakatos, 1970), by explaining more aspects dealing with the comparison of scientific theories. Levi's theory of suppositional reasoning is also related to the problem of so-called "belief change" (Levi, 1996).

> expected. Assimilating a new sort of fact demands a more than additive adjustment of theory, and until that adjustment is completed - until the scientist has learned to see nature in a different way - the new fact is not quite a scientific fact at all (Kuhn, 1970, p. 53).

It is well-known that the recent falsificationist tradition in epistemology has focused attention on the role of anomalies (that can give rise to falsifications) establishing a sort of "received view" on the growth of scientific knowledge characterized by the fundamental role played by anomalies: Newton's theory is able to explain phenomena not touched on by Aristotle's theory, such as correlations between the tides and the location of the moon, and the variation in the force of gravity with respect to height above sea level; in turn, Einstein was able to do the same with respect to the Newtonian theory and its anomalies and falsifications.

As Lakatos argues, in a mature theory with a history of useful consequences, it is generally better to reject an anomalous conflicting report than it is to abandon the theory as a whole. The cases in which we have to abandon a whole theory are very rare: a theory may be considered as a complex information system in which there is a collection of cooperating individual statements some of which are useful and more firmly held than others; propositions that belong to the central core of a theory are more firmly held than those which are located closer to the border, where instead rival hypotheses may coexist as mutually incompatible alternatives. Accumulating reports of empirical observations can help in deciding in favor of one alternative over another.

I have to remember that even without restoring consistency, an inconsistent system can still produce useful information. Of course from the point of view of classical logic we are compelled to derive any conclusion from inconsistent premises, but in practice efficient proof procedures infer only "relevant'' conclusions with varying degrees of accessibility, as stated by the criteria of non-classical *relevant entailment* (Anderson and Belnap, 1975).

We may conclude by asserting that contradiction, far from damaging a system, helps to indicate regions in which it can be changed (and improved). Contradiction has a preference for strong hypotheses which are more easily falsified than weak ones; and moreover, hard hypotheses may more easily weakened than weak ones, which prove difficult subsequently to strengthen. It is always better to produce mistakes and then correct them than to make no progress at all[3].

[3] We have to remember that not all the configurations of new concepts are incoherence-driven and related to the highest case of creative abduction and creative analogical reasoning, ubiquitous in science, and more constructive than associative. For example, in conceptual combination in everyday reasoning, many new concepts are formed in a coherence-driven way, where a kind of reconciliation of associations and thematic relations operates. Tha-

2. GOVERNING INCONSISTENCIES IN ABDUCTIVE REASONING

We can see abductive inferences "as answers to inquirer's explicit (or usually) tacit questions put to some definite source of answers (information)" (Hintikka, 1998, p. 519) stressing the interrogative features of this kind of reasoning. If abduction is the making of a set of possible answers, the choice of the possible questions is also decisive (and this choice of course is not indifferent as regards the further process of finding answers). At present, the cognitive sciences lack a good theory of how fruitful questions are produced, able to involve also emotional, ethical, and pragmatic aspects of cognition (Thagard and Croft, 1999).

I think that the role of *inconsistencies* in scientific reasoning could be very important to delineate the first features of such a theory. Surely surprise and curiosity, are related to the detection of inconsistencies. Model-based ways of generating a hypothesis that explains some phenomenon or conceptual problem that produced the question are heuristically linked to the activity itself of "finding" that certain puzzling phenomenon or that particular conceptual problem. In turn, the deductive component and the inductive testing of hypothesis of the whole cycle of reasoning illustrated above – abduction as inference to the best explanation – clearly shows the presence of a further interrogative feature (further generation of questions).

If we want to deal with the nomological and most interesting creative aspects of abduction we are first of all compelled to consider the whole field of the growth of scientific knowledge cited above. We have anticipated (see chapter 2) that abduction has to be an inference permitting the derivations of *new* hypotheses and beliefs. Some explanations consist of certain facts (initial conditions) and universal generalizations (that is scientific laws) that deductively entail a given fact (observation), as showed by Hempel in his *law covering model* of scientific explanation (Hempel, 1966). If T is a theory illustrating the background knowledge (a scientific or common sense *theory*) the sentence α explains the fact (observation) β just when $\alpha \cup T \models \beta$; we have already illustrated in chapter 2 that it is difficult to treat nomological and causal aspects of abduction and explanation in the framework of the belief revision: we would have to deal with a kind of belief revision that permits us to alter a theory with new conditionals.

---

gard presents the case of the construction of the concept of "computational philosopher" where in order to understand the concept people need to make coherent sense of how a "modifier" such as "computational" can apply to a "head" such as "philosopher" (Thagard, 1997b).

As already illustrated in chapter 2 (section 2)[4] we may see belief change from the point of view of *conceptual change*, considering concepts either cognitively, like mental structures analogous to data structures in computers, or, epistemologically, like abstractions or representations that presuppose questions of justification. Belief revision - even if extended by formal accounts such as illustrated above[5] - is able to represent cases of conceptual change such as adding a new instance, adding a new weak rule, adding a new strong rule (see Thagard, 1992, pp. 34-39, for details), that is, cases of addition and deletion of beliefs, but fails to take into account cases such as adding a new part-relation, adding a new kind-relation, adding a new concept, collapsing part of a kind-hierarchy, reorganizing hierarchies by branch jumping and tree switching, in which there are reorganizations of concepts or redefinitions of the nature of a hierarchy. These last cases are the most evident changes occurring in many kinds of creative reasoning, for example in science. Related to some of these types of conceptual change are different varieties of inconsistencies (see Figure 1), as explained in the following sections.

2.1 Finding inconsistencies I: empirical anomalies

In chapter 2 I argued that logical accounts of abduction certainly illustrate much of what is important in abductive reasoning, especially the objective of selecting a set of hypotheses (diagnoses, causes) that are able to dispense good (preferred) explanations of data (observations), but fail in accounting for many cases of explanations occurring in science or in everyday reasoning. For example they do not capture 1. the role of statistical explanations, where what is explained follows only probabilistically and not deductively from the laws and other tools that do the explaining; 2. the sufficient conditions for explanation; 3. the fact that sometimes the explanations consist of the application of schemas that fit a phenomenon into a pattern without realizing a deductive inference; 4. the idea of the existence of high-level kinds of *creative* abductions; 5. the existence of model-based abductions (for instance visual and diagrammatic); 6. the fact that explanations usually are not complete but only furnish *partial* accounts of the pertinent evidence (see Thagard and Shelley, 1998); 7. the fact that one of the most important virtues of a new scientific hypothesis (or of a scientific theory) is its power of explaining *new*, previously *unknown* facts.

[4] Cf. also Magnani, 1999d.

[5] Or developed by others, see for example, Katsuno and Mendelzon, 1992, Cross and Thomason, 1992.

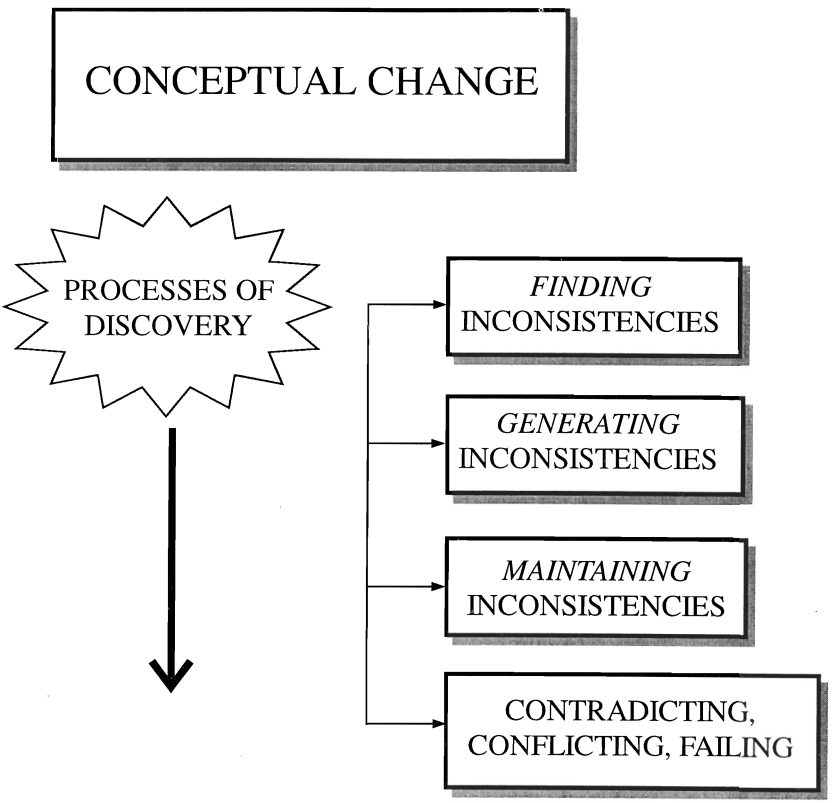

Figure 1. Conceptual change and inconsistencies.

Moreover, the logical accounts of abduction certainly elucidate many kinds of inconsistency government, which nevertheless reduce to the act of finding contradictions able to generate the withdrawal of some hypotheses, beliefs, reasons, etc.: these contradictions always emerge at the level of data (observations), and consistency is restored at the theoretical level[6]. This view may distract from important aspects of other kinds of reasoning that involve intelligent abductive performances.

For example, *empirical anomalies* result from data that cannot currently be fully explained by a theory. They often derive from predictions that fail, which implies some element of incorrectness in the theory. In general terms, many theoretical constituents may be involved in accounting for a given do-

[6] We have to remember that the logical models in some cases exhibit a sort of paraconsistent behavior.

main item (anomaly) and hence they are potential points for modification. The detection of these points involves defining which theoretical constituents are employed in the production of the anomaly. Thus, the problem is to investigate all the relationships in the explanatory area. In science, first and foremost, empirical anomaly resolution involves the localization of the problem at hand within one or more constituents of the theory. It is then necessary to produce one or more new hypotheses to account for the anomaly, and finally, these hypotheses need to be evaluated so as to establish which one best satisfies the criteria for theory justification. Hence, anomalies require a change in the theory, yet once the change is successfully made, anomalies are no longer anomalous but in fact are now resolved. This can involve both transforming the domain knowledge and learning or discovering new schemas or rules endowed with explanatory power. Also in many machine discovery programs (cf. chapter 2, section 5, this book) failed predictions drive the mechanism which selects new experiments to guess new hypotheses (Zytkow, 1997). The process is goal-driven. Of course an explainer will use different information for the objective of predicting a situation than for repairing or preventing it with a good explanation.

General strategies for anomaly resolution, as well as for producing new ideas and for assessing theories, have been studied by Darden (1991) in her book on reasoning strategies from Mendelian genetics. Anomaly resolution presents four aspects: 1. confirmation that the anomaly exists, 2. localization of the problem (not considering the cases where the anomaly either is an uninteresting monster or is outside the scope of the theory), 3. generation of one or more new hypotheses to account for the anomaly, that is, conceptual changes in the theory, 4. evaluation and assessment of the hypotheses chosen. Abductive steps are present at the third level (that normally is activated when step 2 fails to manage the anomaly): in this case we are trying to eliminate the already confirmed anomaly (step 1) in a creative way. At this level there are various kinds of conceptual changes: from the simple ones related to the possible alteration (deletion, generalization, simplification, complication, "slight" changes, proposal of the opposite, etc.), or addition of a new component of the theory, to the construction and discovery of a new "theoretical component" (Darden, 1991, pp. 269-275).

Let us relate this taxonomy to the cases of scientific conceptual change illustrated above. Darden's alterations and additions can be assimilated to the cases of conceptual change such as adding a new instance, adding a new weak rule, adding a new strong rule (that is, cases of addition and deletion of beliefs), adding a new part-relation, adding a new kind-relation, adding a new concept. On the contrary, Darden's case about the discovery of a new "theoretical component" relates to changes where there is collapse of part of a kind-hierarchy, the reorganization of hierarchies by branch jumping and

tree switching, in which there are reorganizations of concepts or redefinitions of the nature of a hierarchy. We have seen that these last cases are the most evident changes occurring in many kinds of creative reasoning in science, when adopting a new conceptual system is more complex than mere belief revision: different varieties of *model-based abductions* are related to some of these types of scientific conceptual change.

Darden reminds to us that geneticists sometimes abandoned hypotheses on the basis of falsifying evidence (this is another way - unsuccessful resolution, unresolved anomaly - of seeing the problem of anomalies, already indicated at the step "alter a component-deletion"). And, of course, as clearly illustrated by Lakatos (see section 1, above) an anomalous observation statement can be rejected and the theory which it clashes retained. This is for example the case of the Copernicus' theory that was retained and the naked-eye observation of the sizes of Venus and Mars, which were in contradiction with that theory, eliminated.

Moreover, as taught by the recent epistemological tradition, and especially by the so-called sophisticated falsificationism (Lakatos, 1970), in science new creative abduced hypotheses (or new theories), originating from a case of anomaly resolution, have to lead to novel predictions. Galileo reported that the moon was not a smooth sphere but full of craters and mountains. His Aristotelian opponent had to admit that observation that nevertheless configured a terrible anomaly for the notion, common to many Aristotelian and coming from the ancient times, that the celestial bodies are perfect spheres. The hypothesis created by the Aristotelian to explain this anomaly is *ad hoc*: he advocated that there is an invisible substance on the moon, which fills the craters and covers the mountains so that the moon's aspect is absolutely spherical. Unfortunately, this modified theory of the moon did not lead to new testable consequences, and thus is not scientifically acceptable, following the falsificationist point of view[7].

2.2 Finding inconsistencies II: conceptual anomalies

Empirical anomalies are not alone in generating impasses, there are also the so-called *conceptual anomalies*. To illustrate more features of a theory of the role of inconsistencies in model-based abduction we can present the case of conceptual problems as triggers for hypotheses. The so-called conceptual problems represent a particular form of anomaly. In addition, resolving conceptual problems may involve satisfactorily answering questions about the nature of theoretical entities. Nevertheless such conceptual problems do not arise directly from data, but from the nature of the claims in the principles or

[7] On the relationship between falsificationism (Popperian and Lakatosian) and conventionalism (at least in the ingenious case of Poincaré), cf. chapter 7, section 1.3.

in the hypotheses of the theory. It is far from simple to identify a conceptual problem that requires a resolution, since, for example, a conceptual problem concerns the adequacy or the ambiguity of a theory, and yet also its incompleteness or (lack of) evidence.

The formal sciences are especially concerned with conceptual problems. Let's consider an example deriving from the well-known case of the non-Euclidean revolution, which plays a remarkable role in illustrating some actual transformations in rational conceptual systems. The discovery of non-Euclidean geometries involves some interesting cases of *visual abductive reasoning*. It demonstrates a kind of visual abduction, as a strategy for anomaly resolution related to a form of explanatory and productive visual thinking.

Since ancient times the fifth postulate has been held to be not evident. This "conceptual problem" (just an anomaly) has caused much suspicion about the reliability of the whole theory of parallels, consisting of the theorems that can be only derived with the help of the fifth postulate. The recognition of this anomaly was fundamental to the development of the great non-Euclidean revolution. Two thousand years of attempts to resolve the anomaly have generated many more-or-less fallacious demonstrations of the fifth postulate (for example, a typical attempt was that of trying to prove the fifth postulate from the others), until the discovery of non-Euclidean geometries (Greenberg, 1980).

I will present in chapter 7, section 2.2 (see also Magnani, 1999b) some details derived from the historical discovery of non-Euclidean geometries which illustrate the relationships between strategies for anomaly resolution and visual thinking: I consider how Lobachevsky's strategy for resolving the anomaly of the fifth postulate was to manipulate the symbols, rebuild the principles, and then to derive new proofs and provide a new mathematical apparatus. The failure of the demonstrations of his predecessors induced Lobachevsky to believe that the difficulties that had to be overcome were due to causes other than those which had until then been focused on. I will show how some of the hypotheses created by Lobachevsky were mostly image-based trying to demonstrate that visual abduction is relevant to hypothesis formation and scientific discovery.

The fact that inconsistencies may occur also at the theoretical level is further emphasized if we consider that in science or in legal reasoning (Thagard and Shelley, 1998), hypotheses are mainly *layered*, contrarily to the case of diagnostic reasoning, where we have a set of data that can be explained by a given set of diseases (that is with the explanation consisting of a mapping from the latter to the former). Hence, the organization of hypotheses is more complex than the one illustrated in formal models, and abduction is not only a matter of mapping from sets of hypotheses to a set of data.

In many abductive settings there are hypotheses that explain other hypotheses so that the selection or creation of explanations is related to these relationships[8]. In this case the plausibility of the hypothesis comes not only from what it explains, but also from it itself being explained. The Darwinian hypothesis stating that "Species of organic beings have evolved" gains plausibility from the many pieces of evidence it helps to explain. Moreover, it receives plausibility from above, from being explained by the hypothesis of natural selection, in its turn explained by the hypothesis concerning the struggle for existence. The principle of special relativity and the principle of the constancy of the speed of light explain (in this case the explanatory relation is "deductive") the Lorentz transformation, which explains the negative result of the Michelson-Morley experiment, but also they explain the convertibility of mass and energy which explains the nuclear transmutations detected by Rutherford in 1919. Hence the two principles explain the two experiments above by means of the intermediate layered hypotheses of Lorentz transformation and mass/energy conversion, but we also know the two principles directly explain the Fizeau experiment concerning the speed of light in a flowing fluid (Einstein, 1961).

In some machine discovery programs the question of layered hypotheses could be related to the one of postulating hidden structures where some hidden hypotheses can trigger discovery of other hypotheses at a higher level.

2.3 Generating inconsistencies by radical innovation

The case of conceptual change such as adding a new part-relation, adding a new kind-relation, adding a new concept, collapsing part of a kind-hierarchy, reorganizing hierarchies by branch jumping and tree switching, in which there are reorganizations of concepts or redefinitions of the nature of a hierarchy are the most evident changes occurring in many kinds of *creative abduction*, for instance in the growth of scientific knowledge.

In *Against Method*, Feyerabend (1993) attributes a great importance to the role of contradiction. He establishes a "counterrule" which is the opposite of the neopositivistic one that it is "experience", or "experimental results" which measures the success of our theories, a rule that constitutes an important part of all theories of corroboration and confirmation. The counterrule "[...] advises us to introduce and elaborate hypotheses which are inconsistent with well-established theories and/or well-established facts. It advises us to proceed counterinductively" (Feyerabend, 1993, p. 20). Counter-induction is seen more reasonable than induction, because appropriate to the needs of creative reasoning in science: "we need a dream-world in order to

[8] This kind of hierarchical explanations has also been studied in the area of probabilistic belief revision (Pearl, 1988).

discover the features of the real world we think we inhabit [...]" (p. 29). We know that counterinduction, that is the act of introducing, inventing, and generating new inconsistencies and anomalies, together with new points of view incommensurable with the old ones, is congruous with the aim of inventing "alternatives" ("proliferation of theories is beneficial for science"), is very important in all kinds of creative abductive reasoning.

When a scientist introduces a new hypothesis, especially in the field of natural sciences, he is interested in the potential rejection of an old theory or of an old knowledge domain. Consistency requirements in the framework of deductive models, governing hypothesis withdrawal in various ways, would arrest further developments of the new abduced hypothesis. In the scientist's case there is not the deletion of the old concepts, but rather the *coexistence* of two rival and competing views.

Consequently we have to consider this competition as a form of epistemological, and non logical inconsistency. For instance two scientific theories are conflicting because they compete in explaining shared evidence.

The problem has been studied in Bayesian terms but also in connectionist ones, using the so-called theory of explanatory coherence (Thagard, 1992, cf. also footnote 2, above), which deals with the epistemological reasons for accepting a whole set of explanatory hypotheses conflicting with another one. In some cognitive settings, such as the task of comparing a set of hypotheses and beliefs incorporated in a scientific theory with the one of a competing theory, we have to consider a very complex set of criteria (to ascertain which composes the best explanation), that goes beyond the mere simplicity or explanatory power. The minimality criteria included in some of the formal accounts of abduction, or the idea of the choice among preferred models cited in section 2 of chapter 2, are not sufficient to illustrate more complicated cognitive situations.

2.4 Maintaining inconsistencies

As noted above, when we create or produce a new concept or belief that competes with another one, we are compelled to maintain the derived inconsistency until the possibility of rejecting one of the two becomes feasible. We cannot simply eliminate a hypothesis and then substitute it with one inconsistent with it, because until the new hypothesis comes in competition with the old one, there is no reason to eliminate the old one. Other cognitive and epistemological situations present a sort of paraconsistent behavior: a typical kind of *inconsistency maintenance* is the well-known case of scientific theories that face anomalies. As noted above, explanations are usually not complete but only furnish partial accounts of the pertinent evidence: not everything has to be explained.

Newtonian mechanics is forced to cohabit with the anomaly of perihelion of Mercury until the development of the theory of relativity, but it also has to stay with its false prediction about the motion of Uranus. In diagnostic reasoning too, it is necessary to make a diagnosis even if many symptoms are not explained or remain mysterious. In this situation we again find the similarity between reasoning in the presence of inconsistencies and reasoning with incomplete information already stressed. Sometimes scientists may generate the so-called auxiliary hypotheses (Lakatos, 1970), justified by the necessity of overcoming these kinds of inconsistencies: it is well-known that the auxiliary hypotheses are more acceptable if able to predict or explain something new (the making of the hypothesis of the existence of another planet, Neptune, was a successful way - not an *ad hoc* maneuver - of eliminating the anomaly of the cited false prediction).

To delineate the first features of a constructive cognitive and formal framework that can handle the coexistence of inconsistent theories (and unify many of the themes concerning the limitations of formal models of abductions previously illustrated) we have first of all to be able to deal with the treatment of non verbal representations (that is model-based representations).

Moreover, I think that the problem of coexistence of inconsistent scientific theories and of reasoning from inconsistencies in scientific creative processes leads to analyze the characters of what I call the *best possible information* of a situation. It is also necessary to distinguish between the dynamic and the static sides of the best possible information. If we stress the *sequential* (dynamic) aspects we are more oriented to analyze anomalies as triggers for hypotheses: as illustrated by the traditional deductive models of abduction, the problem concerns the abductive steps of the sequential comprehension and integration of data into a hypothetical structure that represents the best explanation for them. Analogously, as we will see in the case of conceptual anomaly in geometry (chapter 7, section 2), the "impasse" can also be a trigger for a whole process of model-based abduction. On the contrary, if we consider the *holistic* (static) aspects we are more interested in the coexistence of inconsistencies as potential sources of different reasoned creative processes. In this last case we have to deal with model-based abduction and its possible formal treatment; I plan to derive some suggestions from the area of paraconsistent and adaptive logic (Meheus, 1999), for instance handling hierarchies of inconsistent models of a given representation[9].

When the holistic representation concerns the relationship between two competing theories containing some inconsistencies, a formal framework can be given by the connectionist tradition using a computational reconstruction

[9] See also the analysis of the relationships between inconsistency, generic modeling, and conceptual change given in Nersessian, 1999a.

of the epistemological concept of coherence, as already stated (see also the following section).

2.5 Contradicting, conflicting, failing

Considering the *coherence* of a conceptual system as a matter of the simultaneous satisfaction of a set of positive and negative constraints leads to the *connectionist* models (also in computational terms) of coherence. In this light logical inconsistency becomes a relation that furnishes a *negative* constraint and entailment becomes a relation that provides a *positive* constraint. For example, as already noted, some hypotheses are inconsistent when they simply compete, when there are some pragmatic incompatibility relations, when there are incompatible ways of combining images, etc. (Thagard and Shelley, 1997; Thagard and Verbeurgt, 1998).

From the viewpoint of the connectionist model of coherence, it spontaneously allows the situations in which there is a set of accepted concepts containing an inconsistency, for example in the case of anomalies: the system at hand may at any rate have a maximized coherence, when compared to another system. Moreover, another interesting case is the relation between quantum theory and general relativity, which individually have enormous explanatory coherence. According to the eminent mathematical physicist Edward Witten "the basic problem in modern physics is that these two pillars are incompatible". Quantum theory and general relativity may be incompatible, but it would be folly given their independent evidential support to suppose that one must be rejected (Thagard, 1992, p. 223).

A situation that is specular to inconsistency maintenance (cf. previous section) is given when two theories are not intertranslatable and *observationally equivalent*, as illustrated by the epistemology of conventionalist tradition. In these cases they are unconcerned by inconsistencies (and therefore by crucial experiments, they are unfalsifiable) but have to be seen as rivals. The incommensurability thesis shows interesting relationships with the moderate and extreme conventionalism. If theories that are not intertranslatable, that is incommensurable, function in certain respects as do observationally equivalent theories (and they are unconcerned by crucial experiments), the role of observational and formal-structural invariants in providing comparability is central: it is impossible to find a contradiction in some areas of the conceptual systems they express. I think that it is necessary to study in general the reasons able to model the demise of such observationally equivalent "conventional" theories, showing how they can be motivationally abandoned. This problem has been frequently stressed in the area of automated discovery (cf. chapter 2, section 5): if many hypothetical patterns are discov-

ered, all justified by their observational consequences, we are looking for the reasons to claim that one of them is the best (Zytkow and Fischer, 1996).

Moreover these theories can be seen as rivals in some sense not imagined in traditional philosophy of science. We already stressed that in these cases the role of observational and formal-structural invariants in providing comparability is central: it is impossible to find a contradiction in some area of the conceptual systems they express.

Contradiction has a preference for strong hypotheses which are more easily falsified than weak ones; and moreover, hard hypotheses may more easily weakened than weak ones, which prove difficult subsequently to strengthen: but hypotheses may be *unfalsifiable*. In this case, it is impossible to find a contradiction from the empirical point of view but also from the theoretical point of view, in some area of the conceptual systems in which they are incorporated. Notwithstanding this fact, it is sometimes necessary to construct ways of rejecting the unfalsifiable hypothesis at hand by resorting to some external forms of negation, external because we want to avoid any arbitrary and subjective elimination, which would be rationally or epistemologically unjustified. In the following chapter I will consider a kind of "weak" hypothesis that is hard to negate and the ways for making it easy. I will explore whether *negation as failure* can be employed to model hypothesis withdrawal in Freudian analytic reasoning and in Poincaré's conventionalism of the principles of physics.

2.6 Inconsistencies and narrative abduction

Sentential abduction (chapter 2, section 2) is also active at the level of everyday natural language, when we generate creative or simply new narratives. As in science, restoring coherence is sometimes important in *narratives* as well, many "stories" have to be believable and devoid of emotional and story world inconsistencies, and characters' unmotivated goals.

MINSTREL is a computer program that tells creative stories about King Arthur and his knights (Turner, 1994). To be good, their stories must be consistent at many levels, and the characters and the world should act consistently and predictably. After having created a scene in which Galahad kills a dragon and observing that there is not explanation of why Galahad killed the dragon - this unexplained fact leads to a story inconsistency - MINSTREL adds new story scenes to correct the problem using knowledge about knights and transform and adapt strategies. The program has nine plans for maintaining story consistency and of course it is also endowed with routines which check inconsistencies. Inconsistencies constitute the driving force of the narrative creative abduction. Therefore storytelling can be considered as

a kind of narrative abductive problem solving where the role of inconsistencies is central.

The heuristics that make MINSTREL creative are called Trasform-Recall-Adapt and perform creativity as an integrated process of search and adaptation; moreover, they are integrated with a kind of case-based (or "analogical") reasoning that 1) recalls a paste problem with the same features and its associated solution, 2) adapts and transform - generalization, spacialization, mutation, recombination - the past narrative solution to the current problem, and then 3) assesses the result. One particularly widespread way of creating new story themes is to use generalizations and specializations on existing story themes.

Not only about King Arthur, there are many kinds of narratives, for instance scientists do not use only mathematics, diagrams, and experiments. As illustrated in chapter 3, devoted to the manipulative abduction, scientists use "experimental" narratives in a constructive and hypothesizing role. Moreover, they consent to construct and reconstruct experience, and to distribute it across a social network of negotiations among the different scientists by means of the so-called construals.

In all the narratives, and especially in the narratives of detection, the problem of prompting explanations of actions is widespread (Eco and Sebeok, 1983, Oatley, 1996). We have quoted in chapter 2 (section 3.2) the passage from Peirce about the fact that all sensations participate in the nature of a unifying hypothesis, that is, in abduction, in the case of emotions too: in their cognitive theory of emotions largely based on Peirce's intuitions Oatley and Johnson-Laird (1987) are able to explain how the reader, in front of narratives, feels emotions as abductions.

3. PREINVENTIVE FORMS, DISCONFIRMING EVIDENCE, UNEXPECTED FINDINGS

Intuitively an anomaly is something surprising, as Peirce already knew "The breaking of a belief can only be due to some *novel* experience" (*CP*, 5.524) or "[...] until we find ourselves confronted with some experience contrary to those expectations" (*CP*, 7.36; Peirce, 1955b) (cf. this chapter, section 1)[10]. Therefore it is not strange that something anomalous can be found in those kinds of structures the cognitive psychologists call *preinventive*. Cognitive psychologists have described many kinds of preinventive

[10] Classical cognitive considerations on inconsistencies in reasoning can be found in Schank, 1982 and in Schank and Abelson, 1987.

structures and described their desirable properties, that constitute particularly interesting ways of "irritating" the mind and stimulating creativity.

Preinventive structures are very important from the point of view of creative abduction, because of the propulsive role they play. Finke, Ward, and Smith (1992) (cf. also Finke, 1990) list the following preinventive cognitive structures: *visual patterns* and *objects forms* (one can generate two dimensional patterns resulting in creative products such as new types of symbols and artistic design or three-dimensional forms resulting in new inventions and spatial analogies); *mental blend* (two distinct entities are fused to create something new, one might imagine combining a lion with an ostrich to create a type of animal); exemplars of *unusual* or *hypothetical categories* (they show emergent features that lead to new and unexpected discoveries, for example, in attempting to construct a member of the category "alien creatures that inhabit a planet different from the earth", one might imagine a creature that resemble earth creatures in some respect but not others); *mental models* that represent various mechanical or physical systems (sometimes incomplete, unstable, and even unscientific), as well as conceptual systems; various kinds of *verbal combinations* (they can lead to poetic and other literary and narrative explorations, cf. the previous section). Moreover, some musical forms or actions schemas can be identified, as well as also other possibilities.

Some particular attributes of these structures are very important in contributing to discovery: *novelty*, *ambiguity* (ambiguous visual patterns are often interpreted in various creative ways), *implicit meaningfulness* (they seem to have hidden meanings: "a general perceived sense of 'meaning' in the structure" [...] potential for inspiring or eliciting new and unexpected interpretations", Finke, Ward, and Smith, 1992, p. 23), *emergence* (referred to the extension in the preinventive structures of unexpected relations and features), *incongruity* (that refers to conflict or contrast among elements)[11], *di-*

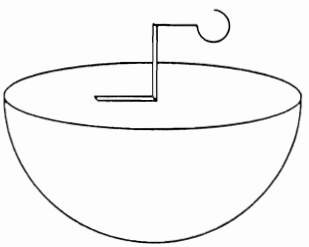

Figure 2. A preinventive form in experiments on spanning the object categories, constructed using the bracket, hook, and half-sphere. (From Finke, Ward, and Smith, 1992, copyright, MIT Press, used by permission).

[11] Already exploited by Koestler's theory of bisociation (1964).

vergence (that refers to the possibility of finding various uses and meanings in the same structure, like in the case of a hammer, an unambiguous form that can be used in many ways). As can be easily seen, all these properties can be considered from the single theoretical point of view of the presence of something anomalous. All properties refer to a kind of detected surprise that can open the abductive exploratory processes of creativity.

Examples of creative reinterpretations of the preinventive form in experiments on spanning the object categories are given in Figures 2 and 3.

Koslowski (1996) studies scientific reasoning observing the principles and strategies people use in generating and testing hypotheses in every day situations. In some situations subjects reason in a scientific way to a greater extent than considered in the existing literature. She illustrates experiments with subjects dealing with *hypothesis-testing* and examines how hypotheses (possible explanations) are generated and how they vary in credibility as a

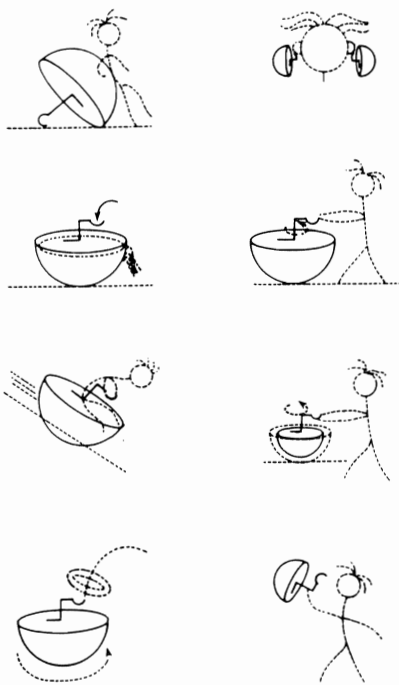

Figure 3. Possible reinterpretations of the preinventive form given in Figure 2, spanning eight object categories (left to right): lawn lounger (furniture), global earrings (personal items), water weigher, (scientific instruments) portable agitator (appliances), water sled (transportation), rotating masher (tool and utensils), ring spinner (toys and games), and slasher basher (weapons). (From Finke, Ward, and Smith, 1992, copyright, MIT Press, used by permission).

function of various sorts of evidence about the considered phenomena. Moreover, the experiments concern the ways in which humans deal with evidence or information "that disconfirms or is anomalous to or at least unanticipated by an explanation" to focus on situations where there is the opportunity to engage hypothesis revision (or hypothesis withdrawal).

The results are quite interesting and show that theoretical concerns play a prominent role in recognizing the importance of the *anomalies*: subjects manage hypothesis revision as theory dependent in a scientific legitimate way; when the alternative explanatory hypotheses involved are in terms of causal (theoretical) mechanisms - and not simply stated in terms of covariation with data - subjects are able to treat them as defeasible and to modify them in ways that are theoretically motivated (Klahr and Dunbar, 1988), not simply using *ad hoc* maneuvers. Decisions about whether to revise or reject as well as decisions about type of revisions that would considered as justified are highly theory-dependent and involve much more that merely information about covariation. In turn, ignoring the importance of the theoretical component (or mechanism), can underestimate subjects' willingness to reject hypotheses when rejection would be appropriate (this is the reason why these subjects have sometimes been regarded as poor scientists, like in the case described in Wason's task, 1960)[12].

Finally, the empirical research of recent years by Dunbar (1995, 1999), in many molecular biology and immunology laboratory in US, Canada and Italy, has demonstrated the central role of the *unexpected* in creative abductive reasoning. Scientists expect the unexpected. By experimentally looking at the so-called "in vivo science" Dunbar analyzes three activities that are seen as the most important in scientific model building: analogical reasoning, attention given to unexpected findings (that is anomalies, errors, inconsistencies), experimental design, and distributed reasoning.

First of all scientists frequently use analogy where there is not a simple answer to a particular problem, and distant analogies are not so widespread as supposed, they are primarily used to explain concepts to others, but not in creative scientific reasoning.

Secondly, it is well know that the recent discoveries of naked DNA and Buckey balls, but also the old well-known of penicillin, nylon, and gravitation are charged to the unexpected: in the "in vivo" science we can see the unexpected is very common, for example it is a regular occurrence that the outcome of an experiment does not match the scientists' prediction. The sci-

[12] On the role of conceptual change in childhood and in "intuitive" theories see Carey (1985, 1996). Analogies and differences between scientific and ordinary thought are illustrated in Kuhn (1991, 1996). Ram, et al. (1995) argue that a creative outcome is not an outcome of extraordinary mental processes, but of mechanisms that are on a continuum with those used in ordinary thinking.

entists have to evaluate which findings are caused by methodological errors, faulty assumptions, and chance events. At the local level of experimentation in real scientific laboratories this research constitutes a kind of confirmation of the Popperian ideas on hypotheses falsification, made in that case at the macro-level of the whole growth of scientific knowledge. The hypotheses are activated to deal with such problematic findings, usually local analogies and model-based abductions, which can give rise to generalizations, causal explanation, visualizations, etc., for finding the common features of the unexpected findings, and possibly discover more general and deep explanations.

Third, experimental design is shown to have interesting cognitive components, illustrating the fact that sometimes the experiments are locally built independently of the hypotheses being tested. The problem is related to the role of manipulative abduction we described in chapter 3, showing how we can find methods of constructivity based on external models and action-based reasoning in scientific and everyday reasoning, like the one embedded in experimental activity. Dunbar says scientists aim firstly at ensuring a robust internal structure of the experiment, optimizing the likelihood experiments will work, performing cost/benefits analyses on possible design components, ensuring acceptance of results in case of negotiation with other scientists of the community involved, and, finally, preferring experiments that have both conditions and control conditions (Dunbar, 1999, p. 95).

Finally, we have to remember that science happens, particularly at the "critical" moments, in a situation of distributed reasoning (see also Thagard, 1997a) by a group of scientists and not individual scientists. Abductive reasoning (to produce multiple hypotheses) and generalization are the main cognitive events that occur during social interactions among scientists. As stressed in chapter 3, real people (and so scientists) are some kinds of cognitive-epistemic "mediating structures" incorporating possible objective cognitive aims: epistemic structures can be embodied in artifacts, in ideas, but also in systems of social interactions.

Chapter 7

Hypothesis Withdrawal in Science

1. WITHDRAWING UNFALSIFIABLE HYPOTHESES

In the previous chapter (section 2.5) I illustrated that contradiction is fundamental in abductive reasoning and that it has a preference for strong hypotheses which are more easily falsified than weak ones. Moreover, hard hypotheses may be more easily weakened than weak ones, which prove difficult subsequently to strengthen. Unfortunately, hypotheses may be *unfalsifiable*. In this case, it is impossible to find a contradiction from the empirical point of view but also from the theoretical point of view, in some area of the related conceptual systems. Notwithstanding this fact, it is sometimes necessary to construct ways of rejecting the unfalsifiable hypothesis at hand by resorting to some external forms of negation, external because we want to avoid any arbitrary and subjective elimination, which would be rationally or epistemologically unjustified.

In this chapter I will consider a kind of "weak" hypothesis in science that is hard to negate and the ways for making it easy. In these cases, the subject can rationally decide to withdraw his hypotheses, and to activate *abductive reasoning,* even in contexts where it is impossible to find "explicit" contradictions; moreover, thanks to the new information reached simply by finding this kind of negation, the subject is free to abduce new hypotheses. I will explore whether *negation as failure* can be employed to model hypothesis withdrawal in Freudian analytic reasoning and in Poincaré's conventionalism of the principles of physics. The first case explains how the questioned

problem of the probative value of clinical findings can be solved, the second one shows how conventions can be motivationally abandoned.

1.1 Negation as failure in query evaluation

Computer and AI scientists have suggested an interesting technique for negating hypotheses and accessing new ones: negation as failure. The objective of this section is to consider how the use of *negation as failure* may be relevant to hypothesis withdrawal. There has been little research into the weak kinds of negating hypotheses, despite abundant reports that hypothesis withdrawal is crucial in everyday life and also in certain kinds of diagnostic or epistemological settings, such as medical reasoning and scientific discovery (cf. chapters 2 and 4, this book).

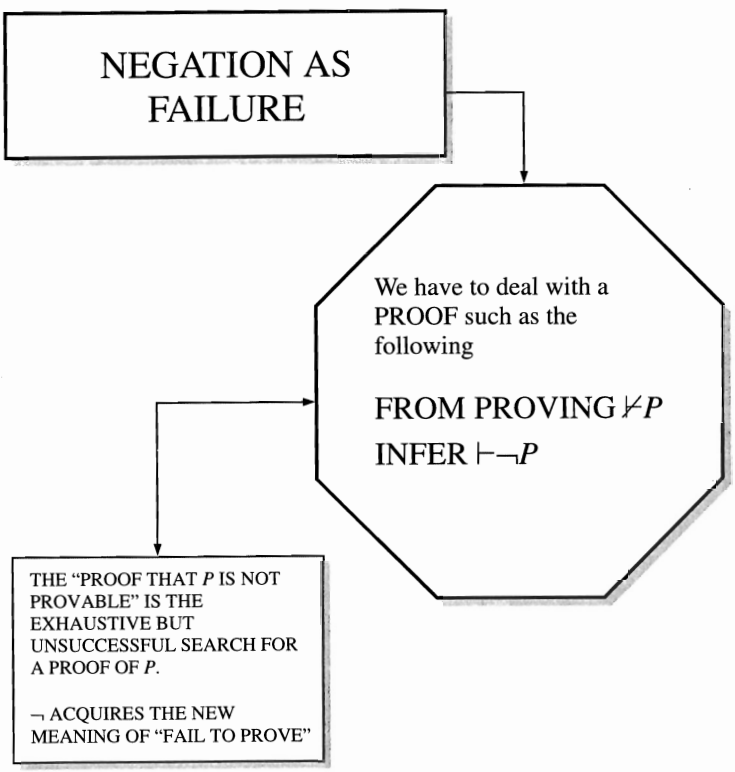

Figure 1. Negation as failure.

In the cases of conceptual change I describe, inferences are made using this kind of negation as a fundamental tool for advancing knowledge: new conclusions are issued on the basis of the data responsible for the failure of the previous ones. I plan to explore whether this kind of negation can be employed to model hypothesis withdrawal in Freudian analytic reasoning and in Poincaré's conventionalism of the principles of physics.

I consider this kind of negation, studied by researchers into logic programming, to be very important also from the epistemological point of view. Negation as failure is active as a "rational" process of withdrawing previously-imagined hypotheses in everyday life, but also in certain subtle kinds of diagnostic (analytic interpretations in psychoanalysis) and epistemological settings. Contrasted with classical negation, with the double negation of intuitionistic logic, and with the philosophical concept of *Aufhebung*[1], negation as failure shows how a subject can decide to withdraw his hypotheses, while maintaining the "rationality" of his argumentations, in contexts where it is impossible to find contradictions.

The statements of a logical database are a set of Horn clauses which take the form:

$$R(t_1, \ldots, t_n) \leftarrow L_1 \wedge L_2 \wedge \ldots \wedge L_m$$

($m \geq 0$, $n \geq 0$, where $R(t_1, \ldots, t_n)$ - conclusion - is the distinguished positive literal[2] and $L_1 \wedge L_2 \wedge \ldots \wedge L_m$ - conditions - are all literals, and each free variable is implicitly universally quantified over the entire implication). In more conventional notation this would be written as the disjunction

$$R(t_1, \ldots, t_n) \vee \neg L_1 \vee \neg L_2 \vee \ldots \vee \neg L_m$$

where any other positive literal of the disjunctive form would appear as a negated precondition of the previous implication.

Let us consider a special query evaluation process for a logical database that involves the so-called *negation as failure* inference rule (Clark, 1978). We can build a Horn clause theorem prover augmented with this special inference rule, such that we are able to infer $\neg P$ when every possible proof of *P fails*.

We know that a relational database only contains information about *true* instances of relations. Even so, many queries involve negation and we can answer them by showing that certain instances are *false*. For example, let's consider this simple case: to answer a request for the name of a student not

[1] Toth (1991) models negation exploiting this Hegelian concept which is very significant for explaining non-Euclidean revolution.

[2] A *literal* is an atomic formula or the negation of an atomic formula.

taking a particular course, C, we need to find a student, S, such that the instance (atomic formula) $Takes(S,C)$ is false. For a logical database, where an atomic formula which is not explicitly given may still be implied by a general rule, the assumption is that an atomic formula is false if we *fail* to prove that it is true. To prove that an atomic formula P is *false* we do an exhaustive search for a proof of P. If *every* possible proof of P fails, we can infer $\neg P$. The well-known PROLOG programming language (Roussel, 1975) uses this method of manipulating negation.

We have to deal with a proof such as the following:

from proving $\nvdash P$ infer $\vdash \neg P$

where the "proof that P is not provable" (Clark, 1978, p. 120) is the *exhaustive* but *unsuccessful search* for a proof of P. Here the logical symbol $\neg$ acquires the new meaning of "fail to prove" (Figure 1).

Clark proposes a query evaluation algorithm based essentially on ordered linear resolution for Horn clauses (SLD) augmented by the negation as failure inference rule "$\neg$ P may be inferred if every possible proof of P fails" (SLDNF)[3].

What is the semantic significance of this kind of negation? Can we interpret a failed proof of $\neg P$ as a *valid* first order inference that P is false? Clark's response resorts to reconciling negation as failure with its truth functional semantics: if we can demonstrate that every failed attempt to prove P using the database of clauses B, is in effect a proof of $\neg P$ using the completed[4] database $C(B)$, then "negation as failure" is a derived inference rule for deductions from $C(B)$: the explicit axioms of equality and completion laws are therefore necessary at the object level in order to simulate failure of the matching algorithm at the meta-level. A negated literal $\neg P$ will be evaluated by recursively entering the algorithmic query evaluator (as an ordered linear resolution proof procedure, as stated above) with the query P. If every possible path for P ends in failure (failure proofs that can be nested to any depth), we return with $\neg P$ evaluated as true.

Clark (1978) has shown that for every meta-language proof of $\neg P$ obtained by a Horn clause theorem prover (query evaluation) augmented with negation as failure there exists a structurally similar object-language proof of $\neg P$. He has proved that a query evaluation with the addition of negation as failure will only produce results that are implied by first order inference from the completed database, that is, the evaluation of a query should be

[3] The links between negation as failure, completed databases (Clark, 1978), and the closed world assumption have been studied in great detail. A survey can be found in (Lloyd, 1987).

[4] The notion of *database completion* can be found in Clark, 1978, and in all textbooks on logic for computer science.

viewed as a "deduction" from the completed database (correctness of query evaluation). Consequently negation as failure is a sound rule for deductions from a completed database.

Although the query evaluation with negation as failure process is in general not complete, its main advantage is the efficiency of its implementation. There are many examples in which the attempt to prove neither succeeds nor fails, because it goes into a loop. To overcome these limitations it is sufficient to impose constraints on the logical database and its queries, and add loop detectors to the Horn clause problem solver: by this method the query evaluation process is guaranteed to find each and every solution to a query. However, because of the undecidability of logic, no query evaluator can identify all cases in which a goal is unsolvable. A best theorem prover does not exist and there are no limitations on the extent to which a problem solver can improve its ability to detect loops and to establish negation as failure.

1.2 Withdrawing constructions

First of all we will illustrate how it is possible to explain the epistemological status of Freud's method of clinical investigation in terms of a special form of negation as failure. I am not dealing here with the highly controversial problem of the epistemological status of psychoanalytic clinical theories (comprehensively analyzed in Grünbaum, 1984): it is well-known that clinical data have no probative value for the confirmation or falsification of the general hypotheses of psychoanalytic clinical theories of personality, because, given that they depend completely on the specific nature of the clinical setting, they are devoid of the independence that characterizes observations endowed with scientific value.

Furthermore, because of the lack of probative value in the patient's clinical data with regard to the analyst's interpretations, any therapeutic gains from analysis may be considered to have been caused not by true insightful self-discovery but rather by placebo effects induced by the analyst's powers of suggestion. If the probative value of the analysand's responses is negated, then Freudian therapy might reasonably be considered to function as an emotional corrective (performed by a positive "transference" effect) and not because it enables the analysand to acquire self-knowledge; instead he or she capitulates to proselytizing *suggestion*, which operates the more insidiously since under the pretense that analysis is nondirective. Suggestion is indeed responsible for the so-called epistemical contamination of the patient's responses.

Freud asks the patient to believe in the analyst's theoretical retrodictions of significant events in his early life and these theoretical retrodictions are communicated to him as *constructions*:

> The analyst finishes a piece of construction and communicates it to the subject of the analysis so that it may work upon it; he then constructs a further piece out of the fresh material pouring in upon him, deals with it in the same way and proceeds in this alternating fashion until the end (Freud, 1953-1974, vol. 23, 1937, pp. 260-261).

The aim is to provoke the previously-cited true insightful self-discovery that guarantees the cure (Freud, cit., 1920, vol. 18, 1920, p. 18). A single construction is built as a "sequence" of the interpretations that issue from clinical data found in the clinical setting, epistemologically characterized by "transference" and "countertransference":

> "Interpretation" applies to something that one does to some single element of the material, such as an association or a parapraxis. But it is a "construction" when one lays before the subject of the analysis a piece of his early history that he has forgotten, in some such way as this: "Up to your *n*th year you regarded yourself as the sole and unlimited possessor of your mother; then came another baby and brought you grave disillusionment. Your mother left you for some time, and even after her reappearance she was never again devoted to you exclusively. Your feelings towards your mother became ambivalent, your father gained a new importance for you," and so on (Freud, cit., vol. 23, 1937, p. 261).

A construction can be considered as a kind of "history" or "narrative" of the analysand's significant early life events, which is never complete, but can be rendered more and more comprehensive by adding new interpretations.

Freudian clinical reasoning refers to a kind of *abductive reasoning*, which I call selective: its uncertainty is due to *nonmonotonicity* (cf. chapter 2, section 1.2, this book), the analyst may always withdraw his or her interpretations (constructions) when new evidence arises. Every construction is generated by a "double" abduction: first of all the analyst has to select a suitable general psychoanalytic hypothesis, apply it to some "single element of the material" to produce an interpretation, then he/she has to select each of these general hypotheses in such a way that the sequence of the generated interpretations can give rise to a significant and consistent construction. Every "abduced" construction, suitably connected with some other clinical psychoanalytical hypotheses, generates expectations with regard to the analysand's subsequent responses and remarks.

Let us remember that Habermas considers therapy as due to a sort of Hegelian causality of fate: the analyst applies what Habermas calls "general

interpretations"[5] (Habermas, 1971, p. 259) to the analysand's clinical data. This application generates particular interpretations that combine into a "narrative" (Freud's "construction"). Within the scientophobic framework of Habermas's philosophy this application is regarded as "hermeneutic", because the constructions are presumed to be expressed in the "intentional" and motivational language of desires, affects, fantasies, sensations, memories, etc.

Of course the analyst aims to build *the most complete* construction. The problem here is the analyst cannot propose to the analysand any construction he wants, without some form of external testing. As stated above, the objection most often raised against psychoanalysis is that "therapeutic success is *non*probative because it is achieved *not* by imparting veridical insight but rather by the persuasive suggestion of fanciful pseudoinsights that merely ring verisimilar to the docile patient" (Grünbaum, 1984, p. 138). In one of his last papers, *Constructions in analysis* (Freud, cit., vol. 23, 1937, pp. 257-269), Freud reports that "a certain well-known man of science" had been "at once derogatory and unjust" because

> He said that in giving interpretations to a patient we treat him upon the famous principle of "Heads I win, tails you lose" [In English in the original]. That is to say, if the patient agrees with us, then the interpretation is right, but if he contradicts us, that is only a sign of his resistance, which again shows that we are right. In this way we are always in the right against the poor helpless wretch whom we are analysing, no matter how he may respond to what we put forward (Freud, cit., vol. 23, 1937, p. 257).

Freud looks for a criterion for justifying, in the clinical setting, the abandonment of constructions that have been shown to be inadequate (it is interesting to note that in the cited article Freud emphasizes the provisional role of constructions referring to them also as "hypotheses" or "conjectures"). This is the fundamental epistemological problem of the method of clinical investigation: Freud is clear in saying that therapeutic success will occur only if incorrect analytic constructions, spuriously confirmed by "contaminated" responses from the patient, are discarded in favor of new correct constructions (that are constitutively *provisional*) derived from clinical data not distorted by the patient's compliance with the analyst's communicated expectations.

Freud then proceeds "to give a detailed account of how we are accustomed to arrive at an assessment of the 'Yes' or 'No' [considered as "direct

[5] They correspond to general "schemes" of possible constructions.

evidences"] of our patients during analytic treatment - of their expression of agreement or of denial" (p. 257).

Analytic constructions cannot be falsified by dissent from the patient because "it is in fact true that a 'No' from one of our patients is not as a rule enough to make us abandon an interpretation as incorrect" (p. 257). It might seem to Freud that patient dissent from an interpretation can be always discounted as inspired by neurotic resistance. It is only "in some rare cases" that dissent "turns out to be the expression of legitimate dissent" (p. 262). A "patient's 'No' is no evidence of the correctness of a construction, though it is perfectly compatible with it" (p. 263). Rather, a patient's 'No' might be more adequately related to the "incompleteness" of the proposed constructions: "the only safe interpretation of his 'No' is that it points to incompleteness" (p. 263).

Even if a patient's verbal assent may result from genuine recognition that the analyst's construction is true, it may nevertheless be spurious because it derives from neurotic resistance, as already seen in his or her dissent. Assent is "hypocritical" when it serves "to prolong the concealment of a truth that has not been discovered" (p. 262). On the other hand, assent is genuine and not hypocritical when patient's verbal assent will be followed and accompanied by new memories: "The 'Yes' has no value unless it is followed by indirect confirmations, unless the patient, immediately after his 'Yes', produces new memories which complete and extend the construction" (p. 262)

Since "Yes" and "No" do not have any importance to test a construction it is necessary to see other facts, such as "the material" that has "come to light" after having proposed a construction to the patient:

> [...] what in fact occurs [...] is rather that the patient remains as though he were untouched by what has been said and reacts to it with neither a "Yes" nor a "No". This may possibly mean no more than that his reaction is postponed; but if nothing further develops we may conclude that we have made a mistake and we shall admit as much to the patient at some suitable opportunity without sacrificing any of our authority (pp. 261-262).

Let us now analyze this situation from the epistemological point of view: the analyst has to withdraw the construction (hypothesis) when he has failed to prove it. Remember that for a logic database the assumption is that an atomic formula is false if we *fail* to prove that it is true. More precisely: as stated above, every construction, suitably connected with some other clinical psychoanalytical hypotheses, generates expectations with regard to the analysand's subsequent responses and remarks. We consider the fact that we can continuously extend and complete a construction by adding the new (*expected*) material that "has come to light" from the patient as proof of the

construction validity. If the patient does not provide this new "material" which is able to extend the proposed construction, this *failure* leads to the withdrawal of the construction itself. So the "proof that a construction is not provable" is the *unsuccessful search* for a proof of the construction itself. Here the logical symbol ¬ acquires the new meaning of "fail to prove" in the empirical sense.

Let us resume: if the patient does not provide new "material" which extends the proposed construction, "if", as Freud declares, "nothing further develops we may conclude that we have made a mistake and we shall admit as much to the patient at some suitable opportunity without sacrificing any of our authority". The "opportunity" of rejecting the proposed construction "will arise" just

> [...] when some new material has come to light which allows us to make a better construction and so to correct our error. In this way the false construction drops out, as if it has never been made; and indeed, we often get an impression as though, to borrow the words of Polonius, our bait of falsehood had taken a carp of truth (p. 262).

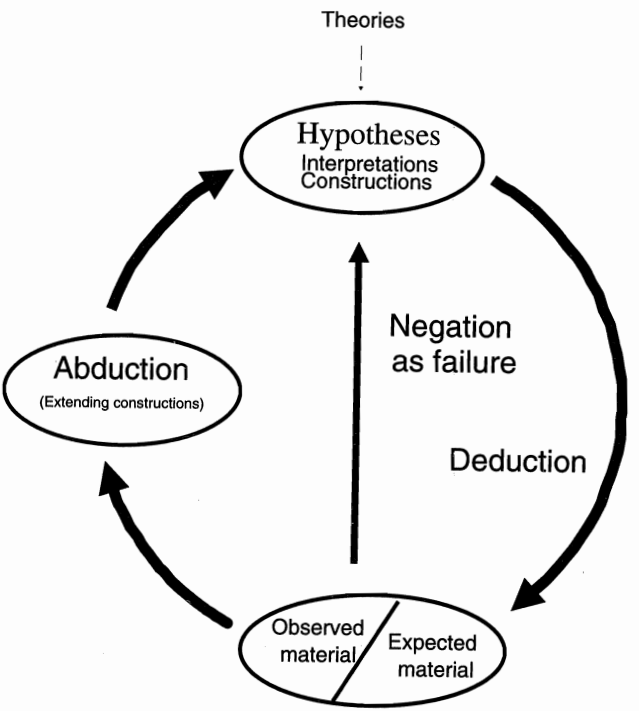

Figure 2. Withdrawing constructions.

A new cycle very similar to the one previously started with the assumption of the first construction takes place: a new construction (derived by applying new clinical psychoanalytical hypotheses and schemes) is provisionally conceived on the basis of the new material that came to light when the analyst was seeking to extend the old one ("we often get an impression as though [...] our bait of falsehood had taken a carp of truth") (cf. Figure 2).

The inferential process is clearly *nonmonotonic*: in an initial phase we have some material coming from the patient and which provides the background for an initial abduced construction; in a second phase we have to add to the initial which emerges after having communicated to the patient the first construction. If in this second phase the new material is not suitable for extending the first construction, the negation as failure compels the analyst to withdraw and reject the construction. The whole process is nonmonotonic because the increase of material does not generate an increase in (the number of) constructions: the old construction is abandoned ("the false construction drops out, as if it has never been made").

I should stress that the epistemological role of what Freud calls "*indirect confirmations*" or *disconfirmations* of analytic constructions is in my opinion negligible. These patient responses, other than verbal assent or dissent, Freud declares, "are in every respect trustworthy" (p. 263). Examples are when a patient has a mental association whose content is similar to that of the construction, or when a patient commits a parapraxis as part of a direct denial. Moreover, when a masochistic patient is averse to "receiving help from the analyst", an incorrect construction will not affect his symptoms, but a correct one will produce "an unmistakable aggravation of his symptoms and of his general condition" (p. 265). The indirect confirmations, in Freud's opinion, provide a "valuable basis for judging whether the construction is likely to be [further] confirmed in the course of analysis" (p. 264).

We should not confuse the kind of transformations of "No" in "Yes" (or vice versa) that pertains to the "indirect confirmations" or disconfirmations, with the above-described extension of constructions toward the most complete one on the basis of new (expected) material that emerges. In conclusion, a patient's "Yes" or "No", whether direct or indirect, has no role to play in withdrawing constructions. I think that these cases do not have any function in the process of abandoning a construction because they always keep open the possibility of extending it: moreover, the indirect confirmations or disconfirmations do not increase the acceptability of constructions. In my opinion, we should not consider Freud as an inductivist, despite his emphasis on these kind of indirect evidence.

Moreover, as stated by Grünbaum, who tends to consider Freudian analysis of clinical method as inductive, this presumption of the consilience of clinical inductions is "spurious" because

> [...] the *independence* of the inferentially concurring pieces of evidence is grievously jeopardized by a *shared* contaminant: the analyst's influence. For *each* of *seemingly* independent clinical data may well be more or less alike confounded by the analyst's suggestion as to conform to his construction, at the cost of their epistemic reliability or probative value. For example a "confirming" early memory may be compliantly produced by the patient on the heels of giving docile assent to an interpretation (Grünbaum, 1984, p. 277).

The second section of *Constructions in analysis* concludes with a very explicit affirmation of nonmonotonicity:

> Only the further course of the analysis enables us to decide whether our constructions are correct or unserviceable. [...] an individual construction is anything more than a conjecture which awaits examination, confirmation or rejection. We claim no authority for it, we require no direct agreement from the patient, nor do we argue with him if at first he denies it. In short, we conduct ourselves on the model of a familiar figure in one of Nestroy's farces - the manservant who has a single answer on his lips to every question or objection: "It will all become clear in the course of future developments" (p. 265).

But I have shown that Freud considers important only the rejection achieved by negation as failure. The epistemological aim is not to validate a construction by extensions provided by new material or by indirect confirmations or disconfirmations. Freud aims to reject it. Perhaps in Freud's considerations there are some ambiguities in perceiving the asymmetry between falsification and confirmation, but it would seem that my interpretation of Freud as a special falsificationist can be maintained without fear of distorting his methodological intention. Freud is a special type of "falsificationist" because negation as failure guarantees the possibility of freely withdrawing a construction and substituting it with a rival and better one. In the computational case, negation as failure is achieved by suitable algorithms related to the knowledge that is handled (see above, section 2). In the human and not computational case, negation as failure is played out in the midst of the analyst-analysand interaction, where transference and countertransference are the human epistemological operators and "reagents". Negation as failure is therefore a limitation on the dogmatic and autosuggestive exaggerations of (pathological) countertransference.

1.3 Withdrawing conventions

We will now consider some aspects dealing with Poincaré's famous conventionalism of the principles of physics and the possibility of negating conventions. An extension of Poincaré's so-called *geometric conventionalism*, according to which the choice of a geometry is only justifiable by considerations of simplicity, in a psychological and pragmatic sense ("commodisme"), is the *generalized conventionalism*, expressing the conventional character of the principles of physics:

> The principles of mathematical physics (for example, the principle of conservation of energy, Hamilton's principle in geometrical optics and in dynamics, etc.) systematize experimental results usually achieved on the basis of two (or more) rival theories, such as the emission and the undulation theory of light, or Fresnel's and Neumann's wave theories, or Fresnel's optics and Maxwell's electromagnetic theory, etc. They express the common empirical content as well as (at least part of) the mathematical structure of such rival theories and, therefore, can (but need not) be given alternative theoretical interpretations (Giedymin, 1982, pp. 27-28).

From the epistemological point of view it is important to stress that the conventional principles usually survive the demise of theories and are therefore responsible for the continuity of scientific progress. Moreover, they are not empirically falsifiable; as stated by Poincaré in *Science and Hypothesis*:

> The principles of mechanics are therefore presented to us under two different aspects. On the one hand, they are truths founded on experiment, and verified approximately as far as almost isolated systems are concerned; on the other hand they are postulates applicable to the whole of the universe and regarded as rigorously true. If these postulates possess a generality and a certainty which the experimental truths from which they were deduced lack, it is because they reduce in final analysis to a simple convention that we have a right to make, because we are certain beforehand that no experiment can contradict it. This convention, however, is not absolutely arbitrary; it is not the child of our caprice. We admit it because certain experiments have shown us that it will be convenient, and thus is explained how experiment has built up the principles of mechanics, and why, moreover, it cannot reverse them (Poincaré, 1902, pp. 135-136).

The conventional principles of mechanics derive from experience, as regards their "genesis", but cannot be falsified by experience because they contribute to "constitute" the experience itself, in a proper Kantian sense. The experience has only suggested their adoption because they are *conven-*

ient: there is a precise analogy with the well-known case of geometrical conventions, but also many differences, which pertain the "objects" studied[6].

Poincaré seeks also to stress that geometry is more abstract than physics, as is revealed by the following speculations about the difficulty of "tracing artificial frontiers between the sciences":

> Let it not be said that I am thus tracing artificial frontiers between the sciences; that I am separating by a barrier geometry properly so called from the study of solid bodies. I might just as well raise a barrier between experimental mechanics and the conventional mechanics of general principles. Who does not see, in fact, that separating these two sciences we mutilate both, and that what will remain of the conventional mechanics when it is isolated will be but very little, and can in no way be compared with that grand body of doctrine which is called geometry (Poincaré, 1902, pp. 137-138).

I believe that the meaning of this passage refers primarily to the fact that physics cannot be considered completely conventional because we know that the conventional "principles" are derived from the "experimental laws" of "experimental mechanics", and then absolutized by the "mind". Second, Poincaré wants to demonstrate how geometry is more abstract than physics: geometry does not require a rich experimental reference as physics does, geometry only requires that experience regarding its genesis and as far as demonstrating that it is the most convenient is concerned. Here we are very close to Kant's famous passage about the *synthetical a priori* character of the judgments of (Euclidean) geometry, and of the whole of mathematics: "The science of mathematics presents the most splendid example of the extension of the sphere of pure reason without the help of the experience" (Kant, 1929, A712-B740, p. 576).

[6] The conventional principles of mechanics should not be confused with geometrical conventions: "The experiments which have led us to adopt as more convenient the fundamental conventions of mechanics refer to bodies which have nothing in common with those that are studied by geometry. They refer to the properties of solid bodies and to the propagation of light in a straight line. These are mechanical, optical experiments" (Poincaré, 1902, pp. 136-137), they are not, Poincaré immediately declares, "*des expériences de géométrie*" (*ibid.*): "And even the probable reason why our geometry seems convenient to us is, that our bodies, our hands, and our limbs enjoy the properties of solid bodies. Our fundamental experiments are pre-eminently physiological experiments which refer, not to the space which is the object that geometry must study, but to our body - that is to say, to the instrument which we use for that study. On the other hand, the fundamental conventions of mechanics and experiments which prove to us that they are convenient, certainly refer to the same objects or to analogous objects. Conventional and general principles are the natural and direct generalisations of experimental and particular principles" (*ibid.*).

Even when separated from the reference to solid bodies, Euclidean geometry maintains all its conceptual pregnancy, as a convention that, in a proper Kantian sense, "constitutes" the ideal solid bodies themselves.

This is not the case of the conventional principles of mechanics when separated from experimental mechanics: "what will remain of the conventional mechanics [...] will be very little" if compared "with that grand body of doctrine which is called geometry".

Poincaré continues:

> Principles are conventions and definitions in disguise. They are, however, derived from experimental laws, and these laws have, so to speak, been erected into principles to which our mind attributes an absolute value. Some philosophers have generalized far too much. They have thought that the principles were the whole of science, and therefore that the whole of science was conventional. This paradoxical doctrine, which is called nominalism, cannot stand examination. How can a law become a principle? (Poincaré, 1902, p. 138).

If the experimental laws of experimental physics are the source of the conventional principles themselves, conventionalism escapes nominalism.

As stated at the beginning of this section, conventional principles survive the demise (falsification) of theories in such a way that they underlie the incessant spectacle of scientific revolutions: "It is the mathematical physics of our fathers which has familiarized us little by little with these various principles; which has habituated us to recognize them under the different vestments in which they disguise themselves" (Poincaré, 1905, p. 95). Underlying revolutions of physics, conventional principles guarantee the historicity and the growth of science itself. Moreover, the conventional principles surely imply "*firstly*, that there has been a *growing tendency* in modern physics to *formulate and solve* physical problems *within powerful, and more abstract, mathematical systems of assumptions* [...]; *secondly*, the role of conventional principles has been growing and *our ability to discriminate experimentally between alternative abstract systems* which, with a great approximation, save the phenomena *has been diminishing* (by comparison to the testing of simple conjunctions of empirical generalizations)" (Giedymin, 1982, p. 28).

Moreover, as stated above, they are not empirically falsifiable: "The principles of mechanics [...] reduce in final analysis to a simple convention that we have a right to make, because we are certain beforehand that no experiment can contradict it" (Poincaré, 1902, p. 136).

Up to now I have considered in detail how the conventional principles guarantee the revolutionary changes of physics and why they cannot be considered arbitrary, being motivated by the *experimental laws* of the "experi-

mental physics", that is by experience. Although arbitrary and conventional, the conventional principles too can be substituted by others. This is the main problem treated by Poincaré in the last passages of Chapter IX, "The Future of Mathematical Physics", in *The Value of Science*. Already the simple case of "linguistic" changes in science "suffices to reveal generalizations not before suspected" (Poincaré, 1905, p. 78). By means of the new discoveries, scientists arrive at a point where they are able to "admire the delicate harmony of numbers and forms; they marvel when a new discovery opens to them an unexpected perspective" (Poincaré, 1905, p. 76), a new perspective that is always provisional, fallible, open to further confirmations or falsifications when compared to rival perspectives.

We have seen how the conventional principles of physics guarantee this continuous extension of experience thanks to the various perspectives and forms expressed by experimental physics. However, because conventional, "no experiment can contradict them". The experience only suggested the principles, and they, since absolute, have become constitutive just of the empirical horizon common to rival experimental theories.

Poincaré observes:

> Have you not written, you might say if you wished to seek a quarrel with me - have you not written that the principles, though of experimental origin, are now unassailable by experiment because they have become conventions? And now you have just told us that the most recent conquests of experiment put these principles in danger. Well, formerly I was right and today I am not wrong. Formerly I was right, and what is now happening is a new proof of it (Poincaré, 1905, p. 109).

Poincaré appeals to a form of weak negation, just as Freud did when dealing with the problem of withdrawing constructions. Let us follow the text. To pursue his point, Poincaré illustrates the attempts to reconcile the "calorimetric experiment of Curie" with the "principle of conservation of energy":

> This has been attempted in many ways; but there is among them one I should like you to notice; this is not the explanation which tends today to prevail, but it is one of those which have been proposed. It has been conjectured that radium was only an intermediary, that it only stored radiations of unknown nature which flashed through space in every direction, traversing all bodies, save radium, without being altered by this passage and without exercising any action upon them. Radium alone took from them a little of their energy and afterward gave it out to us in various forms (Poincaré, 1905, pp. 109-110).

At this point Poincaré resolutely asserts: "What an advantageous explanation, and how convenient! First, it is unverifiable and thus irrefutable. Then again it will serve to account for any derogation whatever to Mayer's principle; it answers in advance not only the objection of Curie, but all the objections that future experimenters might accumulate. This new and unknown energy would serve for everything" (p. 110). Now Poincaré can show how this *ad hoc* hypothesis can be identified with the non-falsifiability of the conventional principle of the conservation of energy:

> This is just what I said, and therewith we are shown that our principle is unassailable by experiment. But then, what have we gained by this stroke? The principle is intact, but thenceforth of what use is it? It enabled us to foresee that in such and such circumstance we could count on such total quantity of energy; it limited us; but now that this indefinite provision of new energy is placed at our disposal, we are no longer limited by anything (Poincaré, 1905, p. 110).

Finally, Poincaré's argumentation ends by affirming negation as failure: "and, as I have written in 'Science and Hypothesis', if a principle ceases to be fecund, experiment without contradicting it directly will nevertheless have condemned it" (*ibid.*) (cf. Figure 3).

Let us now analyze this situation from the epistemological point of view: the conventional principle has to be withdrawn when it "ceases to be fecund", or when it seems that we have failed to prove it. Remember that for a logic database the assumption is that an atomic formula is false if we *fail* to prove that it is true. More clearly: as stated above, every conventional principle, suitably underlying some experimental laws, generates *expectations* with regard to the subsequent evidences of nature. We consider as proof of a conventional principle the fact that we can increasingly *extend* and complete the experimental laws related to it, adding the new (expected) evidence that "emerges" from the experimental research. If, after a finite period of time, nature does not provide this new "evidence" that is able to increase the fecundity of the conventional principle, this *failure* leads to its withdrawal: "experiment without contradicting it directly will nevertheless have condemned it". Analogously to the Freudian case of constructions I have illustrate in the previous section the "proof that a principle is not provable" is the unsuccessful search for a proof of the principle itself. Here too, the logical symbol $\neg$ acquires the new meaning of "fail to prove" in the empirical sense.

Let us resume: if the old conventional principle does not produce new experimental "evidence" to underpin it, it is legitimate to abandon the principle, when convenient: the opportunity to reject the old principle will happen just by exploiting the experimental evidence which, even if not suitable for contradicting it (that is, it is "unassailable by experiment"), is neverthe-

less suitable as a basis for conceiving a new alternative principle, generated by new *creative abductions*.

We can now interpret Popper's ideas about conventionalism in a different way. Popper writes: "Thus, according to the conventionalist view, it is not possible to divide systems of theories into falsifiable and non-falsifiable ones; or rather, such a distinction could be ambiguous. As a consequence, our criterion of falsifiability must turn out to be useless as a criterion of demarcation" (Popper, 1959, p. 81). In the light of Poincaré's theory of the principles of physics that we have just illustrated, the nominalistic interpretation of conventionalism given by Popper (see also Popper, 1963) appears to be very reductive. Moreover, Popper's tendency to identify conventions with *ad hoc* hypotheses (a very bad kind of auxiliary hypotheses) is shown to be decidedly unilateral, since, as is demonstrated by the passages, immediately above, the *adhocness* is achieved only in a very special case, when the conventional principle is epistemologically exhausted.

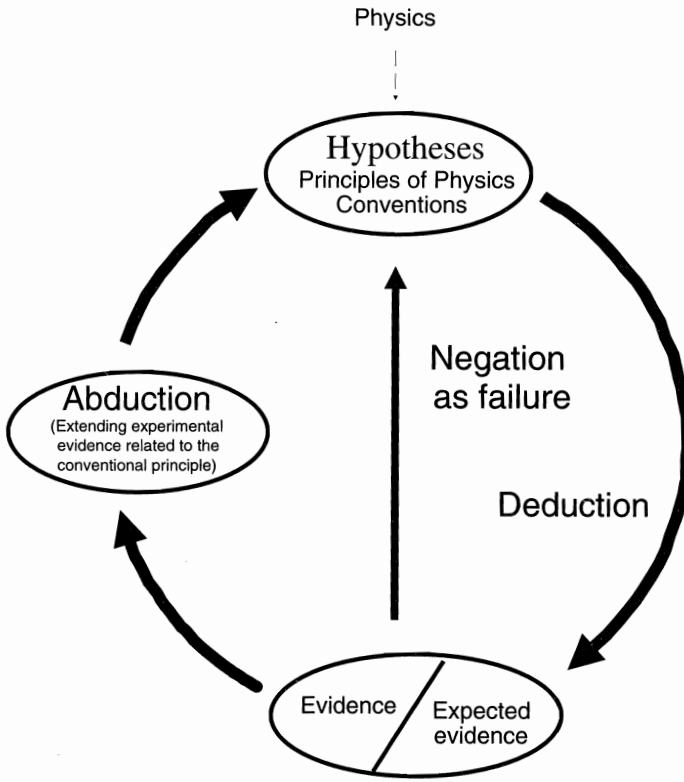

Figure 3. Withdrawing conventions.

In some sense Poincaré was already aware of the following fact, subsequently clearly acknowledged by Popper: the introduction of auxiliary hypotheses must not diminish the degree of falsifiability or testability (that will have to be performed by means of "severe tests") of the scientific theory in question, but, on the contrary, should increase it (Popper, 1959, p. 83). The example of an unsatisfactory (*ad hoc*) auxiliary hypothesis given by Popper is the "[...] contraction hypothesis of Fitzgerald and Lorentz which had no falsifiable consequences but merely served to restore the agreement between theory and experiment - mainly the findings of Michelson and Morley" (Popper, 1959, p. 83).

In turn sophisticated falsificationism (Lakatos, 1970) has definitely established that modified hypotheses (by means of auxiliary assumptions) have to be more falsifiable than the original versions, they have to lead to new testable consequences (and, moreover, independently testable, to use Popper's phrase - 1959, p. 193); progress in "scientific programs" is heavily related to the existence of novel predictions: one program is superior to another insofar as it is a more successful predictor of novel phenomena.

Something analogous operates in the case of the conventional principles described by Poincaré: it seems that conventionalism, at least in the Poincaré's case, does not treat all hypotheses like "stratagems", as maintained by Popper. Hypothetical conventional principles are unfalsifiable and should be withdrawn only when exhausted, when their indirect production of "novel" evidence is finished. Consequently, Poincaré's conventionalism is not simply a theory of *adhocness*, in the nominalistic Popperian sense.

2. THEORETICAL ANOMALY RESOLUTION

2.1 Scientific concept formation and spatial thinking

Among the different abductive roles played by various kinds of conceptual transformations in developing scientific creative reasoning, such as anomaly resolution, conceptual combination, analogical and visual thinking, thought experiment, etc., this section considers how the use of visual/spatial thinking may be relevant to the creation of scientific hypotheses. From this perspective, some examples deriving from the historical discovery of non-Euclidean geometries will be presented in order to demonstrate the relationships between strategies for anomaly resolution and explanatory and productive visual thinking

As previously illustrated in chapters 1 and 5, visual/imagery (but also analogical) reasoning can be productive in scientific concept formation too,

where the role they play in model-based abductive reasoning is very manifest. We said that visual abduction, but also many kinds of abductions involving analogies, diagrams, thought experimenting, visual imagery, etc. in scientific discovery processes, can be just called *model-based.*

How does this kind of analogical and/or imagery reasoning function in scientific problem-solving? Nersessian (1984, 1995a and b, 1998, 1999b) has demonstrated that history of science abounds with instances of the use of imagery and of analogy to transform vague notions into scientifically viable conceptualizations of a domain. Her analysis deals with the important case of the use of imagery and analogy by Faraday and Maxwell in the construction of the concept of field. The concept of field had its origins in vague speculations about processes in the regions surrounding bodies and charges that might contribute to their action upon one another. In articulating a field representation for electric and magnetic actions, Faraday used primarily qualitative concepts and reasoned from imagery figures. He created a field representation for electric and magnetic actions by reasoning from an imagery representation of the "lines of force" that are formed when iron filings are sprinkled around a magnetic source. Many features of "lines" are incorporated into his field concept. He discussed many actions as "expanding", "bending", "being cut". All the forces of nature are unified and interconvertible through various motions of the lines of forces, and matter, itself, is nothing but point centers of converging lines of force. This representation enabled Faraday to express a quantitative relationship between the number of lines cut and the intensity of the induced force. At the end of this research Faraday introduced a pictorial representation that played an important role in Maxwell's construction of the quantitative field concept.

Use of analogy, imagery, and visual/spatial thinking in ordinary and scientific problem-solving is very complex. Nevertheless, we may observe in many cases all the features of a *productive*, *creative* mapping, where such "transfer of knowledge" is essential to the development of a new concept. Imagery representations appear to function analogically. The value of an imagery representation is that it makes some structural relations immediately evident[7].

2.2 Anomaly resolution and spatial reasoning

Empirical anomalies result from data that cannot currently be fully explained by a theory. They often derive from predictions that fail, which implies some element of incorrectness in the theory. In general terms, many theoretical constituents may be involved in accounting for a given domain

[7] The so-called *constructive* and *generic* modeling in scientific discovery are illustrated in Nersessian, et al. (1997).

item (anomaly) and hence they are potential points for modification (chapter 6, section 2.1). The detection of these points involves defining which theoretical constituents are employed in the explanation of the anomaly. Thus, the problem is to investigate all the relationships in the explanatory area.

As illustrated in the previous chapter, first and foremost, anomaly resolution involves the localization of the problem at hand within one or more constituents of the theory, it is then necessary to produce one or more new hypotheses to account for the anomaly, and, finally, these hypotheses need to be evaluated so as to establish which one best satisfies the criteria for theory justification. Hence, anomalies require a change in the theory. We know that empirical anomalies are not alone in generating impasses. The so-called *conceptual problems* represent a particular form of anomaly (cf. chapter 6, section 2.2). Resolving conceptual problems may involve satisfactorily answering questions about the nature of theoretical entities. Conceptual problems do not arise directly from data, but from the nature of the claims in the principles or in the hypotheses of the theory. Usually it is necessary to identify the conceptual problem that needs a resolution, for example by delineating how it can concern the adequacy or the ambiguity of a theory, and yet also its incompleteness or (lack of) evidence.

The formal sciences are especially concerned with conceptual problems. The discovery of non-Euclidean geometries involves an interesting case of visual/spatial abductive reasoning. It demonstrates a kind of *visual/spatial abduction*, as a strategy for anomaly resolution related to a form of explanatory and productive visual thinking. Since ancient times the fifth postulate has been held to be not evident. This "conceptual problem" has caused many difficulties about the reliability of the theory of parallels, consisting of the theorems that can be only derived with the help of the fifth postulate. The recognition of this anomaly was crucial to the development of the great non-Euclidean revolution. Two thousand years of attempts to resolve the anomaly have generated many fallacious demonstrations of the fifth postulate: a typical attempt was that of trying to prove the fifth postulate from the others. Nevertheless, these attempts have also provided much theoretical speculation about the unicity of Euclidean geometry and about the status of its principles.

Here, we are primarily interested in showing how the anomaly is recognizable. A postulate that is equivalent to the fifth postulate states that for every line l and every point P that does not lie on l, there exists a unique line m through P that is parallel to l. This may seem "evident" to the reader, but this is because we have been conditioned to think in terms of Euclidean geometry. The definition represents the level at which ancient Euclidean geometry was developed as a formal science - a level composed of *symbols* and *propositions*.

We can immediately detect a difference between this postulate and the other four if we regard the first principles of geometry as abstractions from experience that we can in turn represent by drawing figures on a blackboard or on a sheet of paper or on our "visual buffer" (Kosslyn and Koenig, 1992) in the mind. We have consequently a *double passage* from the sensorial experience to the abstraction (expressed by symbols and propositions) and from this abstraction to the experience (sensorial and/or mental).

We immediately discover that the first two postulates are abstractions from our experiences drawing with a straightedge, the third postulate derives from our experiences drawing with a compass. The fourth postulate is less evident as an abstraction, nevertheless it derives from our measuring angles with a protractor (where the sum of supplementary angles is 180°, so that if supplementary angles are congruent to each other, they must each measure 90°) (Greenberg 1980, p. 17).

In the case of the fifth postulate we are faced with the following serious problems: 1) we cannot verify empirically whether two lines meet, since we can draw only segments, not lines. Extending the segments further and further to find if they meet is not useful, and in fact we cannot continue indefinitely. We are forced to verify parallels indirectly, by using criteria other than the definition; 2) the same holds with regard to the representation in the "limited" visual buffer. The "experience" localizes a problem to solve, an ambiguity, only in the fifth case: in the first four cases our "experience" *verifies* without difficulty the abstraction (propositional and symbolic) itself. In the fifth case the formed images (mental or not) are the images that are able to *explain* the "concept" expressed by the definition of the fifth postulate as problematic (an anomaly): we cannot draw or "imagine" the two lines at infinity, since we can draw and imagine only segments, not the lines themselves.

The *selected* visual/spatial image or imagery derived from the propositional and symbolic level of the definition is nothing more than the explanation of the anomaly of the definition itself. As stated above, the image demonstrates a kind of visual abduction, as a strategy for anomaly localization related to a form of explanatory visual/spatial thinking, very similar to that illustrated in chapter 5 (section 1), when the case of detecting the anomaly of the presence/absence of an object was described. The only important difference relates to the fact that in that case we dealt only with relationships between images, while here we have to consider relationships between propositional-symbolic levels and visual-imagery ones.

Once the anomaly is detected, the way to anomaly resolution is opened up - in our case, this means that it becomes possible to discover non-Euclidean geometries. That Euclid himself did not fully trust the fifth postulate is revealed by the fact that he postponed using it in a proof for as long as

possible - until the twenty-ninth proposition. As is well-known, Proclus tried to solve the anomaly by proving the fifth postulate from the other four. If we were able to prove the postulate in this way, it would become a theorem in a geometry which does not require that postulate (a future "absolute geometry") and which would contain all of Euclid's geometry.

Without showing all the passages of Proclus's argument (Greenberg 1980, pp. 119-121) we need only remember that the argument seemed correct because it was proved using a diagram. Yet we are not allowed to use that diagram to justify a step in a proof. Each step must be proved from stated axioms or previously proven theorems. We may visualize parallel lines as railroad tracks, everywhere equidistant from each other, and the ties of the tracks as being perpendicular to both parallels. Yet this imagery is valid only in Euclidean geometry. In the absence of the parallel postulate we can only consider two lines as "parallel" when, by the definition of "parallel", they do not possess any points in common. It is not possible implicitly to assume that they are equidistant; nor can it be assumed that they have a common perpendicular. This is an example in which a *selected* abduced image is capable of compelling you to make a mistake, and in this way it was used as a means of evaluation in a proof: we have already stated that it is not possible to use that image or imagery to justify a step in a proof and, moreover, it is not possible to use that image or imagery that attributes to experience more than the experience itself can deliver.

For over two thousand years some of the greatest mathematicians tried to prove Euclid's fifth postulate. For example, Saccheri's strategy for anomaly resolution in the 18th century was to propose two opposites (Darden, 1991, pp. 272-275) of the principle, that is, to negate the fifth postulate and derive, using new logical tools coming from non-geometrical sources of knowledge, all theorems from the two alternative hypotheses by trying to detect a contradiction. The aim was indeed that of demonstrating that the anomaly is simply apparent.

The contradiction in the elliptic case ("hypothesis of obtuse angle", to use the Saccheri's term designing one of the two elementary non-Euclidean geometries) was found, but the contradiction in the hyperbolic case ("hypothesis of the acute angle") was not so easily discovered: having derived several conclusions that are now well-known propositions of non-Euclidean geometry, Saccheri was forced to resort to a metaphysical strategy for anomaly resolution: "Proposition XXXIII. The 'hypothesis' of acute angle [that is, the hyperbolic case] is absolutely false, because repugnant to the nature of the straight line" (Saccheri, 1920). Saccheri chose to state this result with the help of the somewhat complicated imagery of infinitely distant points: two different straight lines cannot both meet another line perpendicularly at one point, if it is true that all right angles are equal (fourth postulate) and that

two different straight lines cannot have a common segment. Saccheri did not ask himself whether everything that is true of ordinary points is necessarily true of an infinitely distant point. In Note II to proposition XXI some "physico-geometrical" experiments to confirm the fifth postulate are also given, invalidated unfortunately by the same incorrect use of imagery that we have observed in Proclus's case. In this way, the anomaly was resolved unsatisfactorily and Euclid was not freed of every fleck: nevertheless, although he did not recognize it, Saccheri had discovered many of the propositions of non-Euclidean geometry (Torretti, 1978, p. 48).

Geometers were not content merely to manipulate proofs in order to discover new theorems and thereby to resolve the anomaly without trying to answer questions about what the symbols of the principles underlying Euclidean geometry represent. Let me illustrate an example where we can see the abductive role played in a discovery process by new considerations concerning *visual sense impressions* and *productive imagery representations*.

Lobachevsky's strategy for resolving the anomaly of the fifth postulate was to manipulate the symbols, rebuild the principles, and then to derive new proofs and provide a new mathematical apparatus. Needless to say, Lobachevsky was working in a specific cultural and scientific environment and his analysis depended on the previous mathematical attempts to demonstrate the fifth postulate. The failure of the demonstrations of his predecessors induced Lobachevsky to believe that the difficulties that had to be overcome were due to causes other than those which had until then been focused on.

Lobachevsky was obliged first of all to rebuild the basic principles: to this end, it was necessary to consider geometrical principles in a new way, as neither ideal nor *a priori*. New interrelations were created between two areas of knowledge: Euclidean geometry and the philosophical tradition of empiricism/sensualism. From this perspective the abductive explanation of the basic concepts of any science is in terms of senses: the basic concepts are always acquired through our *sense impressions*. Geometry is built upon the concepts of body and bodily contact, the latter being the only "property" common to all bodies that we ought to call geometrical. The well-known concepts of depthless surface, widthless line and dimensionless point were constructed considering different possible kinds of bodily contact and avoiding, *per abstractionem*, everything but contact itself: these concepts "exist only in our representation; whereas we actually measure surfaces and lines by means of bodies" for "in nature there are neither straight lines nor curved lines, neither plane nor curved surfaces; we find in it only bodies, so that all the rest is created by our imagination and exists just in the realm of theory" (Lobachevsky, 1897, p. 16). The only thing that we can know in nature is movement "without which sense impressions are impossible. Conse-

quently all other concepts, e.g. geometrical concepts, are generated artificially by our understanding, which derives them from the properties of movement; this is why space in itself and by itself does not exist for us" (Lobachevsky, 1897, p. 9).

This leads Lobachevsky to abduce a very remarkable and modern hypothesis, which I consider to be largely *image-based:* since geometry is not based on a perception of space, but constructs a concept of space from an experience of bodily movement produced by physical forces, there could be place in science for two or more geometries, governing different kinds of natural forces:

> To explain this idea, we assume that [...] attractive forces decrease because their effect is diffused upon a spherical surface. In ordinary Geometry the area of a spherical surface of radius r is equal to $4\pi r^2$, so that the force must be inversely proportional to the square of the distance. In Imaginary Geometry I found that the surface of the sphere is
>
> $$\pi\,(e^r - e^{-r})^2,$$
>
> and it could be that molecular forces have to follow that geometry [...]. After all, given this example, merely hypothetical, we will have to confirm it, finding other more convincing proofs. Nevertheless we cannot have any doubts about this: forces by themselves generate everything: movement, velocity, time, mass, matter, even distances and angles (Lobachevsky, 1897, p. 9).

Lobachevsky did not doubt that something, not yet observable with a microscope or analyzable with astronomical techniques, accounted for the reliability of the new non-Euclidean imaginary geometry. Moreover, the principles of geometry are held to be testable and it is possible to prepare an experiment to test the validity of the fifth postulate or of the new non-Euclidean geometry, the so-called *imaginary geometry*. He found that the defect of the triangle formed by Sirius, Rigel and Star No. 29 of Eridanus was equal to $3.727 + 10^{-6}$ seconds of arcs, a magnitude too small to be significant as a confirmation of imaginary geometry, given the range of observational error. Gauss too had claimed that the new geometry might be true on an astronomical scale. Lobachevsky says:

> Until now, it is well-known that, in Geometry, the theory of parallels had been incomplete. The fruitlessness of the attempts made, since Euclid's time, for the space of two thousand years, aroused in me the suspicion that the truth, which it was desired to prove, was not contained in the data themselves; that to establish it the aid of experiment would be needed, for example, of astronomical observations, as in the case of other laws of nature. When I had finally convinced myself of the justice of my conjec-

> ture and believed that I had completely solved this difficult question I wrote, in 1826, a memoir on this subject *Exposition succincte des principes de la Géométrie* (Lobachevsky, 1897, p. 5).

With the help of the abductive role played by the new sensualist considerations of the basic principles, by the empiricist view and by a very remarkable productive visual hypothesis, Lobachevsky had the possibility to proceed in discovering the new theorems. Following Lobachevsky's discovery the fifth postulate will no longer be considered in any way anomalous - we do not possess any proofs of the postulate, because this proof is impossible. Moreover, the new non-Euclidean hypothesis is reliable: indeed, to understand visual thinking we have also to capture its status of guaranteeing the reliability of a hypothesis.

In order to prove the relative consistency of the new non-Euclidean geometries we should consider some very interesting visual and mathematical "models" (i.e. the Beltrami-Klein and Poincaré models), which involve new uses of visual images in theory assessment.

References

Abe, A., 2000, On the relation between abductive and inductive hypotheses, in: P. Flach and A. Kakas, eds., pp. 169-180.

Ackermann, R., 1989. The new experimentalism, *British Journal for the Philosophy of Science* 40:185-90.

Ajjanagadde, V., 1991, Abductive reasoning in connectionist networks, Technical Report, Wilhelm Schickard Institute, Tübingen, Germany.

Akins, K., ed., 1996, *Perception*, Oxford University Press, Oxford, 1996.

Alchourrón, C., Gärdenfors, P., and Makinson, P., 1985, On the theory of logic change: partial meet functions for contractions and revision, *Journal of Symbolic Logic* 50: 510-530.

Aliseda, A., 1997, *Seeking Explanations: Abduction in Logic, Philosophy of Science, and Artificial Intelligence*, Ph.D. Thesis, Institute for Logic, Language, and Computation (ILLC), University of Amsterdam, The Netherlands.

Aliseda, A., 2000, Abduction as epistemic change: a Peircian model in artificial intelligence, in: P. Flach and A. Kakas, eds., pp. 45-58.

Allen, J.F., 1984, Towards a general theory of action and time, *Artificial Intelligence* 23:123-154.

Allwein, G. and Barwise, J., eds., 1996, *Logical Reasoning with Diagrams*, Oxford University Press, New York.

Anderson, D.R., 1986, The evolution of Peirce's concept of abduction, *Transactions of the Charles S. Peirce Society* 22(2):145-164.

Anderson, D.R., 1987, *Creativity and the Philosophy of Charles Sanders Peirce*, Clarendon Press, Oxford.

Anderson, A. and Belnap, N., 1975, *Entailment*, Princeton University Press, Princeton.

Anderson, J.R. and Libiere, C., 1998, *The Atomic Components of Thought*, Erlbaum, Hillsdale, NJ.

Aravindan, C. and Dung, P.M., 1995, Knowledge base dynamics, abduction, and databases updates, *Journal of Applied Non-Classical Logics* 5:51-76.

Ayim, M., 1974, Retroduction: the rational instinct, *Transaction of the Charles S. Peirce Society*, 10(1):34-43.

Bacon, F., 1620, *The New Organon*, in: M.R. Matthews, ed., 1989, pp. 47-52.

Bailer-Jones, D.M., 1999, Tracing the development of models in the philosophy of science, in: L. Magnani, N.J. Nersessian, and P. Thagard, eds., pp. 23-40.

Barsalou, L.W., 1999, Perceptual symbol systems, *Behavioral and Brain Sciences* 22:577-609.

Barwise, J. and Etchemendy, J., 1991, Visual information and valid reasoning, in: *Visualization in Teaching and Learning Mathematics*, W. Zimmerman and S. Cunningham, eds., Mathematical Association of America.

Bennett, B., 1994, Spatial reasoning with propositional logic, in: *4th International Conference on Knowledge Representation and Reasoning*, Morgan Kaufmann, Los Altos, CA.

Benthem, J., van, 1983, *The Logic of Time*, Kluwer, Dordrecht.

Bessant, B., 2000, On the relationships between induction and abduction: a logical point of view, in: P. Flach and A. Kakas, eds., pp. 77-88.

Bhatta, S.R. and Goel, A.K., 1997, A functional theory of design patterns, in: *Proceedings of IJCAI-97*, pp. 294-300.

Biot, J.-B., 1821, On the magnetism impressed on metals by electricity in motion; read at the public setting of the Academy of Sciences, 2nd April, 1821, *Quarterly Journal of Science* 11:281-290.

Blades, M., 1997, Research paradigms and methodologies for investigating children's way-finding, in: N. Foreman and R. Gillett, eds., vol. I, pp. 103-129.

Bloch, H. and Morange, F., 1997, Organising gestures in external space: orienting and reaching, in: N. Foreman and R. Gillett, eds., vol. I, pp. 15-40.

Blois, M.S., 1984, *Information and Medicine: The Nature of Medical Description*, University of California Press, Berkeley.

Boden, M., 1991, *The Creative Mind: Myths and Mechanisms*, Basic Books, New York.

Boneh, D., Dunworth, C., Lipton, R.J., and Sgall, J., On the computational power of DNA, *Discrete Applied Mathematics* (Special Issue on *Computational Molecular Biology*) 71: 79-94.

Boshuizen, P.A. and Schmidt, H.G., 1992, On the role of biomedical knowledge in clinical reasoning by experts, intermediates and novices, *Cognitive Science* 16:153-184.

Boutilier, C. and Becher, V., 1995, Abduction as belief revision, *Artificial intelligence*, 77:43-94.

Bovet, J., 1997, Long-distance travels and homing: dispersal, migrations, excursions, in: N. Foreman and R. Gillett, eds., vol. II, pp. 239-269.

Bradie, M., 1974, Polanyi on the Meno paradox, *Philosophy of Science* 41:203.

Brewka, G., 1989, Preferred subtheories: an extended logical framework for default reasoning, *Proceedings IJCAI-89*, Detroit, MI, pp. 1043-1048.

Bringsjord, S., 2000, Is (Gödelian) model-based deductive reasoning computational?, in: Special Issue *Abduction and Scientific Discovery*, L. Magnani, N.J. Nersessian, and P. Thagard, eds., *Philosophica*, forthcoming.

Brown, J.R., 1997, Proofs and pictures, *British Journal for the Philosophy of Science* 48:161-180.

Bruner, J.S., On perceptual readiness, *Psychological Review* 64(2):123-152.

Bruner, J.S., Goodnow, J.J., and Austin, G.A., 1956, *A Study of Thinking*, Wiley, New York.

Buchanan, B.G., 1985, Steps toward mechanizing discovery, in: *Logic of Discovery and Diagnosis in Medicine*, K.F. Schaffner, ed., University of California Press, Berkeley and Los Angeles, pp. 94-114.

Bylander, T., Allemang, D., Tanner, M.C., and Josephson, J.R., 1991, The computational complexity of abduction, *Artificial Intelligence* 49:25-60.

Carey, S., 1985, *Conceptual Change in Childhood*, MIT Press, Cambridge, MA.

Carey, S., 1996, Cognitive domains as modes of thought, in: D.R. Olson and N. Torrance, eds., pp. 187-215.

Carnap, R., 1950, *Logical Foundations of Probability*, Routledge and Kegan Paul, London.

Chalmers, A.F., 1999, *What is this Thing Called Science* (1976), Hackett, Indianapolis/Cambridge.

Chandrasekaran, B., 1983, Towards a taxonomy of problem solving types, *AI Magazine* 4:9-17.

Chandrasekaran, B. and Mittal, S., 1982, Deep versus compiled knowledge in diagnostic problem solving, in: *Proceedings of the National Conference on Artificial Intelligence*, pp. 349-354.

Chandrasekaran, B. and Mittal, S., 1983, Conceptual representation of medical knowledge for diagnosis by computer: MDX and related systems, in: *Advances in Computers*, M. Yovits, ed., Academic Press, New York, pp. 217-293.

Chandrasekaran, B. and Mittal, S., 1986, Generic tasks in knowledge-based reasoning, *IEEE Expert* 1:23-30.

Chandrasekaran, B. and Mittal, S., 1988, Generic tasks as building blocks for knowledge-based systems: the diagnosis and routine design examples, *The Knowledge Engineering Review* 3(3):183-210.

Charniak, E., 1988, Motivation analysis: abductive unification and nonmonotonic equality, *Artificial Intelligence* 34(3):275-295.

Charniak, E. and McDermott, D., 1985, *Introduction to Artificial Intelligence*, Addison-Wesley, Reading, MA.

Châtelet, G., 1999, *Figuring Space. Philosophy, Mathematics, and Physics* (1993), translated by R. Shore and M. Zagha, Kluwer, Dordrecht.

Chi, M.T.H., Feltovich, P.J., and Glaser, R., 1981, Categorization and representation of physics problems by experts and novices, *Cognitive Science* 5:121-52.

Chrisley, R.L., 1995, Taking embodiment seriously: non conceptual content and robotics, in: K.M. Ford, C. Glymour, and P.J. Hayes, eds., pp. 141-166.

Christiansen, H., 2000, Abduction and induction combined in a metalogic framework, in: P. Flach and A. Kakas, eds., pp. 195-212.

Church, A., 1936, A note on the Entscheidungsproblem, *Journal of Symbolic Logic* 1:40-41, Correction, *ibid.*, 101-102.

Clancey, W.J., 1981, NEOMYCIN: reconfiguring a rule-based expert system for application to teaching, in: *Proceedings of the International Conference of Artificial Intelligence*, pp. 829-836.

Clancey, W.J., 1984, Extensions to rules from explanation and tutoring, in: *Rule-Based Expert Systems*, Addison-Wesley, Reading, MA, pp. 351-371.

Clancey, W.J., 1985, Heuristic classification, *Artificial Intelligence* 27:289-350.

Clancey, W.J., 1986, From GUIDON to NEOMYCIN and HERACLES in twenty short lessons (ONR Final Report 1979-1985), *AI Magazine* 7(3):40-60.

Clark, A., 1997, The dynamical challenge, *Cognitive Science* 21(4):461-481.

Clark, K.L., 1978, Negation as failure, in: *Logic and Data Bases*, H. Gallaire and J. Minker, eds., Plenum, New York, pp. 119-140. (Reprinted in: M.L. Ginsberg, ed., 1987, pp. 311-325).

Colton, S., ed., 1999, *AI and Scientific Creativity. Proceedings of the AISB99 Symposium on Scientific Creativity*, Society for the Study of Artificial Intelligence and Simulation of Behaviour, Edinburgh College of Art and Division of Informatics, University of Edinburgh, Edinburgh.

Console, L. and Torasso, P., 1991, A spectrum of logical definitions of model-based diagnosis, *Computational Intelligence* 7(3):133-141. Also in: *Reading in Model-Based Diagnosis*, W. Hamscher, L. Console, and J. de Kleer, eds., Morgan Kaufmann, Los Altos, CA, 1992.

Console, L., Theseider Dupré, D., and Torasso, P., 1991, On the relationship between abduction and deduction, *Journal of Logic and Computation*, 1(5):661-690.

Console, L. and Saitta, L., 2000, On the relations between abductive and inductive explanations, in: P. Flach and A. Kakas, eds., pp. 133-151.

Cornuéjols, A., Tiberghien, A., and Collet, G., 2000, A new mechanism for transfer between conceptual domains in scientific discovery and education, in: Special Issue *Model-Based Reasoning in Scientific Discovery: Learning and Discovery*, L. Magnani, N.J. Nersessian, and P. Thagard, eds., *Foundations of Science*, forthcoming.

Corruble, V. and Ganascia J.-G., 1997, Induction and the discovery of the causes of scurvy: a computational reconstruction, *Artificial Intelligence* 91:205-223.

Coz, P.T. and Pietrsykowski, T., 1986, Causes for events: their computation and application, in: *Proceedings of the 8th CADE*, LNCS 230, Springer, Berlin, pp. 608-621.

Cross, C. and Thomason, R.H., 1992, Conditionals and knowledge-base update, in: Gärdenfors, ed., pp. 247-275.

Darden, L., 1991, *Theory Change in Science: Strategies from Mendelian Genetics*, Oxford University Press, Oxford.

Davies, J. and Goel, A.K., 2000, A computational theory of visual analogical transfer, Technical Report, Georgia Institute of Technology, Atlanta, GA.

Davis, W.H., 1972, *Peirce's Epistemology*, Nijhoff, The Hague.

Davy, H., 1821, On the magnetic phenomena produced by electricity, *Philosophical Transactions* 111:7-19.

Debrok, G., 1997, The artful riddle of abduction (abstract), in: M. Rayo, A. Gimate-Welsh, and P. Pellegrino, eds., p. 230.

Degli Antoni, G. and Pizzi, R., 1991, Virtuality as a basis for problem solving?, *AI & Society* 5:239-254.

de Kleer, J., 1986, An assumption-based TMS, *Artificial Intelligence* 28: 127-162.

de Kleer, J., Mackworth, A.K., and Reiter, R., 1990, Characterizing diagnoses, in: *Proceedings AAAI-90*, Boston, MA, pp. 324-330.

de Kleer, J., Mackworth, A.K., and Reiter, R., 1992, Characterizing diagnoses and systems, *Artificial Intelligence* 56(2-3):197-222. Also in: *Reading in Model-Based Diagnosis*, W. Hamscher, L. Console, and J. de Kleer, eds., Morgan Kaufmann, Los Altos, CA, 1992.

De Raedt, L. and Bruynooghe, M., 1991, A multistrategy interactive concept-learner and theory revision system, in: *Proceedings of the 1st International Workshop on Multistrategy Learning*, Harpers Ferry, pp. 175-190.

Dimopoulos, Y. and Kakas, A.C., 1996, Abduction and inductive learning, in: *Advances in Inductive Logic Programming*, De Raedt, L., ed., IOS Press, Amsterdam, pp. 144-171.

Dorigo, M. and Colombetti, M., 1998, *Robot Shaping. An Experiment in Behavior Engineering*, MIT Press, Cambridge, MA.

Doyle, J., 1979, A truth maintenance system, *Artificial Intelligence* 12:231-272.

Doyle, J., 1989, Constructive belief and rational representation, *Computational Intelligence* 5:1-11.

Doyle, J., 1992, Reason maintenance and belief revision: foundations versus coherence theories, in P. Gärdenfors, ed., pp. 29-51.

Doyle, J., 1998, Artificial intelligence and rational self-government. Technical Report No. CMU-CS-88-124, Computer Science Department, Carnegie Mellon University, Pittsburgh.

Dunbar, K., 1995, How scientists think: online creativity and conceptual change in science, in: *The Nature of Insight*, R.J. Sternberg and J.E. Davidson, eds., MIT Press, Cambridge, MA, pp. 365-395.

Dunbar, K., 1999, How scientists build models in vivo science as a window on the scientific mind, in: L. Magnani, N.J. Nersessian, and P. Thagard, eds., pp. 85-99.

Eco, U. and Sebeok, T.A., 1983, *The Sign of Three. Holmes, Dupin, Peirce*, Indiana University Press, Bloomington, IN.

Einstein, A., 1961, *Relativity: The Special and the General Theory,* Crown, New York.

Einstein, A. and Besso, M., 1972, *Correspondence 1903-1955*, Hermann, Paris.

Eshelman, L., 1989, MOLE: a knowledge-acquisition tool for cover-and-differentiate systems, in: S. Marcus, ed., pp. 37-80.

Eshghi, K., 1988, Abductive planning with event calculus, in: *Proceedings of the 5th International Conference on Logic Programming*, pp. 562-579.

Evans, C., 1989, Negation-as-failure as an approach to the Hanks and McDermott problem, in: *Proceedings of the 6th International Symposium on Artificial Intelligence.*

Evans, J., 1982, *The Psychology of Deductive Reasoning*, Routledge and Kegan Paul, London.

Evans, D.A. and Gadd, C.S., 1989, Managing coherence and context in medical problem-solving discourse, in: *Cognitive Science in Medicine. Biomedical Modeling*, D.A. Evans and V.L. Patel, eds., MIT Press, Cambridge, MA, pp. 211-255.

Fajtlowicz, S., 1988, On conjectures of Graffiti, *Discrete Mathematics* 72:113-118.

Falkenhainer, B.C., 1990, A unified approach to explanation and theory formation, in: J. Shrager and P. Langley, eds., pp. 157-196.

Falkenhainer, B.C., Forbus, D., and Gentner, D., 1990, The structure mapping engine: algorithm and examples, *Artificial Intelligence* 41:1-63.

Fann, K.T., 1970, *Peirce's Theory of Abduction*, Nijhoff, The Hague.

Faraday, M., Historical sketch on electromagnetism, *Annals of Philosophy* 18:195-200, 274-290; 19:107-121.

Farah, M.J., 1988, The neuropsychology of mental imagery: converging evidence from brain-damaged and normal subjects, in: *Spatial Cognition. Brain Bases and Development*, J. Stiles-Davis, M. Kritchevsky and U. Bellugi, eds., Erlbaum, Hillsdale, NJ.

Feltovich, P.J. and Barrows, H.S., 1984, Issues of generality in medical problem-solving, in: *Tutorials in Problem-Based Learning: A New Direction in Teaching the Health Professions*, H.G. Schmidt and M.L. de Volder, eds., Van Gorcum, Assen, Holland.

Fetzer, J.K., 1990, *Artificial Intelligence: Its Scope and Limits*, Kluwer Academic Publishers, Dordrecht.

Feyerabend, P., 1993, *Against Method* (1975), third edition, Verso, London-New York.

Finin, T. and Morris, G., 1989, Abductive reasoning in multiple faults diagnosis, *Artificial Intelligence Review* 3:129-158.

Finke, R.A., 1989, *Principles of Mental Imagery*, MIT Press, Cambridge, MA.

Finke, R.A., 1990, *Creative Imagery: Discoveries and Invention in Visualization*, Erlbaum, Hillsdale, NJ.

Finke, R.A. and Slayton, K., 1988, Explorations of creative visual synthesis in mental imagery, *Memory and Cognition* 16: 52-257.

Finke, R.A., Ward, T.B., and Smith, S.M., 1992, *Creative Cognition. Theory, Research, and Applications*, MIT Press, Cambridge, MA.

Flach, J.M. and Warren, R., 1995, Active psychophysics: the relation between mind and what matters, in: *Global Perspectives on the Ecology of Human-Machine Systems*, J.M. Flach, J. Hancock, P. Caird, and K. Vincente, eds., Erlbaum, Hillsdale, NJ, pp. 189-209.

Flach, P. and Kakas, A., 2000a, Abduction and induction: background and issues, in: P. Flach and A. Kakas, eds., pp. 1-29.

Flach, P. and Kakas, A., eds., 2000, *Abductive and Inductive Reasoning: Essays on Their Relation and Integration*, Kluwer Academic, Dordrecht.

Fodor, J., 1983, *The Modularity of Mind*, MIT Press, Cambridge, MA.

Forbus, K.D., 1984, Qualitative process theory, *Artificial Intelligence* 24:85-168.

Forbus, K.D., 1986, Interpreting measurements of physical systems, *Proceedings of the Fifth National Conference on Artificial Intelligence*, Morgan Kaufmann, Philadelphia, pp. 113-117.

Ford, K.M., Glymour, C., and Hayes, P.J., eds., 1995, *Android Epistemology*, MIT Press, Cambridge, MA.

Foreman, N. and Gillett, R., 1997a, General introduction, in: N. Foreman and R. Gillett, eds, vol. I, pp. 1-14.

Foreman, N. and Gillett, R., eds, 1997, *A Handbook of Spatial Research Paradigms and Methodologies*, Psychology Press, Taylor and Francis, Hove, East Sussex, 2 vols.

Frankfurt, H., 1958, Peirce's notion of abduction, *Journal of Philosophy* 55:594-595

Freud, S., 1916, *Leonardo da Vinci: a Study in Sexuality*, Brill, New York.

Freud, S., 1953-1974, *The Standard Edition of the Complete Psychological Works of Sigmund Freud*, translated by J. Strachey in collaboration with A. Freud, et al., Hogarth Press, London.

Fuller, S., De Mey, M., Shinn, T., and Woolgar, S., 1989, *The Cognitive Turn*, Kluwer, Dordrecht.

Galilei, G., 1610, *The Starry Messenger*, in: *Discoveries and Opinions of Galileo*, translated and edited by S. Drake, pp. 23-58, Doubleday, New York, 1957.

Galilei, G., 1632, *Dialogues Concerning the Two Chief World Systems*, translated by S. Drake, in: M.R. Matthews, ed., 1989, pp. 61-81.

Galilei, G., 1638, *Dialogues and Mathematical Demonstrations Concerning Two New Sciences*, translated by H. Crew and A. de Salvio, MacMillan, New York, 1914.

Gärdenfors, P., 1988, *Knowledge in Flux*, MIT Press, Cambridge.

Gärdenfors, P., ed., 1992, *Belief Revision*, Cambridge University Press, Cambridge.

Gärling, T., Selart, M., and Böök, A., 1997, Investigating spatial choice and navigation in large-scale environments, in: N. Foreman and R. Gillett, eds., vol. I, pp. 153-175.

Gelder, T., van and Port, R.F., 1995, It's about time: an overview of the dynamical approach to cognition, in: R.F. Port and T. van Gelder, eds., pp. 1-44.

Gentner, D., 1982, Are scientific analogies metaphors, in: *Metaphor: Problems and Perspectives*, D.S. Miall, ed., Harvester, Brighton, pp. 107-132.

Gentner, D., 1983, Structure-mapping: a theoretical framework for analogy, *Cognitive Science* 7:155-170.

Gentner, D., Brem, S., Ferguson, R., Wolff, P., Markman, A.B., Forbus, K., 1997, Analogy and creativity in the work of Johannes Kepler, in: *Creative Thought: An Investigation of Conceptual Structures and Processes*, T.B. Ward, S.M. Ward, and J. Vaid, eds., American Psychological Association, Washington, DC, pp. 403-459.

Giaquinto, M., 1994, Epistemology of visual thinking in elementary real analysis, *British Journal for the Philosophy of Science* 45:789-813.

Giedymin, J., 1982, *Science and Convention. Essays on Henri Poincare's Philosophy of Science and the Conventionalist Tradition,* Pergamon, Oxford.

Giere, R.N., 1988, *Explaining Science: A Cognitive Approach*, University of Chicago Press, Chicago.

Giere, R.N., 1999, Using models to represent reality, in: L. Magnani, N.J. Nersessian, and P. Thagard, eds., pp. 41-57.

Ginsberg, M.L., ed., 1987, *Readings in Nonmonotonic Reasoning*, Morgan Kaufman, Los Altos, CA.

Glasgow, J.I., 1993, The imagery debate revisited: a computational perspective, *Computational Intelligence* 9(4):309-333.

Glasgow, J.I. and Papadias, D., 1992, Computational imagery, *Cognitive Science* 16:355-394.

Glymour, C., 1989, When less is more, in: *Cognitive Science in Medicine: Biomedical Modeling*, D.A. Evans and V.L. Patel, eds., MIT Press, Cambridge, MA, pp. 349-367.

Glymour, C., 1992, Thoroughly modern Meno, in: *Inference, Explanation, and Other Frustrations*, J. Earman, ed., University of California Press, Berkeley and Los Angeles, pp. 3-22.

Glymour, C., Scheines, R., Spirtes, P., and Kelly, K., 1987, *Discovering Causal Structure*, Academic Press, San Diego, CA.

Goel, A., (1989), What is abductive reasoning? *Neural Network Review* 3(4):181-187.

Gooding, D., 1990, *Experiment and the Making of Meaning*, Kluwer, Dordrecht.

Gooding, D., 1996, Creative rationality: towards an abductive model of scientific change, *Philosophica* 58:73-102.

Gooding, D. and Addis, T.R., 1999, A simulation of model-based reasoning about disparate phenomena, in: L. Magnani, N.J. Nersessian, and P. Thagard, eds., pp. 103-123.

Gorman, M.E., 1998, *Transforming Nature. Ethics, Invention and Discovery*, Kluwer, Dordrecht.

Graßhoff, G. and May, M., 1995, From historical case studies to systematic methods of discovery, in: *AAAI Symposium on Systematic Methods of Scientific Discovery*, Technical Report SS-95-03, AAAI Press, Menlo Park, CA, pp. 46-57.

Greenberg, M.J., 1980, *Euclidean and Non-Euclidean Geometries*, Freeman and Company, New York (Reprint).

Gregory, R.L., 1987, Perception as hypothesis, in: *The Oxford Companion to the Mind*, R.L. Gregory, ed., Oxford University Press, New York, pp. 608-611.

Groen, G.J. and Patel, V.L., The relationship between comprehension and reasoning in medical expertise, in: *The Nature of Expertise*, M.T.H. Chi, R. Glaser, and M.J. Farr, eds., Erlbaum, Hillsdale, NJ, pp. 287-310.

Grünbaum, A., 1984, *The Foundations of Psychoanalysis. A Philosophical Critique*, University of California Press, Berkeley and Los Angeles, CA.

Habermas, J., 1968, *Erkenntnis und Interesse*, Suhrkamp, Frankfurt am Main, (*Knowledge and Human Interests*, 1971, translated by J. J. Shapiro, Beacon Press, Boston).

Hacking, I., 1983, *Representing and Intervening. Introductory Topics in the Philosophy of Natural Science*, Cambridge University Press, Cambridge.

Hanks, S. and McDermott, D., 1987, Nonmonotonic logic and temporal projection, *Artificial Intelligence* 33:379-384.

Hanson, N.R., 1958, *Patterns of Discovery. An Inquiry into the Conceptual Foundations of Science,* Cambridge University Press, Cambridge.

Harman, G., 1965, The inference to the best explanation, *Philosophical Review* 74:88-95.

Harman, G., 1968, Enumerative induction as inference to the best explanation, *Journal of Philosophy* 65(18):529-533.

Harman, G., 1986, *Change in View. Principles of Reasoning*, MIT Press, Cambridge, MA.

Harris, T., 1999, A hierarchy of models and electron microscopy, in: L. Magnani, N.J. Nersessian, and P. Thagard, eds., pp. 139-148.

Helden, A., van, 1989, Galileo, telescopic astronomy, and the Copernican system, in: *Planetary Astronomy from Renaissance to the Rise of Astrophysics. Part A: Tycho Brahe to Newton*, R. Taton and C. Wilson, eds., Cambridge University Press, Cambridge, pp. 81-105.

Hempel, C.G., 1966, *Philosophy of Natural Science*, Prentice-Hall, Englewood Cliffs, NJ.

Hendricks, F.V. and Faye, J., 1999, Abducting Explanation, in: L. Magnani, N.J. Nersessian, and P. Thagard, eds., pp. 271-294.

Hintikka, J., 1998, What is abduction? The fundamental problem of contemporary epistemology, *Transactions of the Charles S. Peirce Society* 34:503-533.

Hintikka, J. and Remes, U., 1974, *The Method of Analysis. Its Geometrical Origin and Its General Significance*, Reidel, Dordrecht.

Hinton, G., 1979, Some demonstrations of the effects of structural descriptions in mental imagery, *Cognitive Science* 3:231-250.

Hobbs, J., Stickel, M., Appelt, D., and Martin, P., 1993, Interpretation as abduction, *Artificial Intelligence*, 63:60-142.

Holland, K.J., Holyoak, K.J., Nisbett, R.E., and Thagard, P., 1987, *Induction. Processes of Inference, Learning, and Discovery*, Cambridge, MA: MIT Press.

Holton, G., 1972, On trying to understand scientific genius, *American Scholar* 41:95-110.

Holyoak, K.J. and Thagard, P., 1989, Analogical mapping by constraint satisfaction, *Cognitive Science* 13:295-355.

Holyoak, K.J. and Thagard, P., 1995, *Mental Leaps. Analogy in Creative Thought*, MIT Press, Cambridge, MA.

Holyoak, K.J. and Thagard, P., 1997, The analogical mind, *American Psychologist* 52(1):35-44.

Hookway, C., 1992, *Peirce*, Routledge and Kegan Paul, London.

Hunt, E., 1989, Cognitive science: definition, status, and questions, *Annals Review of Psychology* 40:603-629.

Hutchins, E., 1995, *Cognition in the Wild*, MIT Press, Cambridge, MA.

Inoue, K. and Haneda, H., 2000, Learning abductive and nonmonotonic logic programs, in: P. Flach and A. Kakas, eds., pp. 213-231.

Ioerger, T.R., 1994, The manipulation of images to handle indeterminacy in spatial reasoning, *Cognitive Science* 18:551-593.

Ippolito, M.F. and Tweney, R.D., 1995, The inception of insight, in: *The Nature of Insight*, R.J. Sternberg and J.E. Davidson, eds., MIT Press, Cambridge, MA.

Ironi, L., Stefanelli, M., and Lanzola, G., 1992, Qualitative models in medical diagnosis, in: *Deep Models for Medical Knowledge Engineering*, E. Keravnou, ed., Elsevier, Amsterdam, pp. 51-70.

Jackson, P., 1989, Propositional abductive logic, in: *Proceedings of the 7th AISB*, pp. 89-94.

Jackson, P., 1990. Abduction and conterfactuals, in: *Working Notes: AAAI Spring Symposium on Automated Abduction*, Stanford University, pp. 77-81.

Jacobs, R.A. and Kosslyn, S.M., 1994, Encoding shape and spatial relations: the role of receptive field size in coordinating complementary representations, *Cognitive Science* 18:361-386.

Jenkins, M.A., Glasgow, J.I., and McCrosky, C., 1986, Programming styles in NIAL, *IEEE Software* 86:46- 55.

Johnson, L. and Keravnou, E.T., 1988, *Expert Systems Architectures*, Kogan Page, London.

Johnson-Laird, P.N., 1983, *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*, Harvard University Press, Cambridge, MA.

Johnson-Laird, P.N., 1988, *The Computer and the Mind: An Introduction to Cognitive Science*, Harvard University Press, Cambridge, MA.

Johnson-Laird, P.N., 1993, *Human and Machine Thinking*, Erlbaum, Hillsdale, NJ.

Josephson, J.R., 1989a, A layered abduction model of perception: integrating bottom-up and top-down processing in a multi-sense agent, in: *Proceedings of the NASA Conference on Space Telerobotics*, Pasadena, JPL Publications, pp. 89-97.

Josephson, J.R., 1989b, Speech understanding based on layered abduction: working notes for the symposium on spoken language systems, in: *Proceedings of the AAAI-89 Spring Symposium Series*, AAAI Press, Stanford.

Josephson, J.R., 1994a, Conceptual analysis of abduction (second section written with M.C. Tanner), in: J.R. Josephson and S.G. Josephson, eds., pp. 5-29.

Josephson, J.R., 1994b, Perception and language understanding (parts also written with M.C. T. Patten, R. Fox, O. Fujimura, D. Erickson, K. Lenzo), in: J.R. Josephson and S.G. Josephson, eds., 238-261.

Josephson, J.R., 2000a, Smart inductive generalizations are abductions, in: P. Flach and A. Kakas, eds., pp. 31-44.

Josephson, J.R., 2000b, Abduction-prediction model of scientific inference reflected in a prototype system for model-based diagnosis, in: Special Issue *Abduction and Scientific Discovery*, L. Magnani, N.J. Nersessian, and P. Thagard, eds, *Philosophica*, forthcoming.

Josephson, J.R., Chandrasekaran, B., Smith, J.W. jr., and Tanner, M.C., 1986, Abduction by classification and assembly, in: *PSA 1986*, vol. I, Philosophy of Science Association, pp. 458-470.

Josephson, J.R. and Josephson, S.G., eds., 1994, *Abductive Inference. Computation, Philosophy, Technology*, Cambridge University Press, Cambridge.

Kakas, A., Kowalski, R.A., and Toni, F., 1992, Abductive logic programming, *Journal of Logic and Computation* 2(6):719-770.

Kakas, A. and Riguzzi, F., 1997, Learning with abduction, in: *Proceedings of the 7th International Workshop on Inductive Logic Programming*, Lecture Notes in Artificial Intelligence, N. Lavrac and S. Dzeroski., eds., Springer, Berlin, pp. 181-188.

Kant, I., 1929, *Critique of Pure Reason*, translated by N. Kemp Smith, MacMillan, London, reprint 1998; originally published 1787.

Kapitan, T., 1990, In what way is abductive inference creative?, *Transactions of the Charles S. Peirce Society* 26(4)-449-512.

Katsuno, H. and Mendelzon, A., 1992, On the difference between updating a knowledge base and revising it, in: Gärdenfors, P., ed., pp. 183-203.

Kautz, H.A., 1986,. The logic of persistence, in: *Proceedings AAAI 86.*

Kirlik, A., 1998, The ecological expert: acting to create information to guide action, in: *Proceedings of the 1998 Conference on Human Interaction with Complex Systems* (HICS'98), Piscataway, NJ, IEEE Press.

Kirsh, D., 1995, The intelligent use of space, *Artificial Intelligence* 73:31-68.

Klahr, D. and Dunbar, K., 1988, Dual space search during scientific reasoning, *Cognitive Science* 12: 1-48

Koestler, A, 1964, *The Act of Creation*, MacMillan, New York.

Kolodner, J., 1993, *Case-Based Reasoning*, Morgan Kaufmann, San Mateo, CA.

Konolige, K., 1990, Towards a general theory of abduction, in: *Working Notes: AAAI Spring Symposium on Automated Abduction*, Stanford University, pp. 62-66.

Konolige, K., 1992, Abduction versus closure in causal theories, *Artificial Intelligence* 53:255-272.

Koslowski, B., 1996, *Theory and Evidence. The Development of Scientific Reasoning*, MIT Press, Cambridge, MA.

Kosslyn, S.M., 1980, *Image and Mind*, Harvard University Press, Cambridge, MA.

Kosslyn, S.M., 1983, *Ghosts in the Mind's Machine: Creating and Using Images in the Brain*, W.W. Norton, New York.

Kosslyn, S.M. and Koenig, O., 1992, *Wet Mind, the New Cognitive Neuroscience*, Free Press, New York.

Krauß, S., Martignon, L., and Hoffrage, U., 1999, Simplifying Bayesian inference: the general case, in: L. Magnani, N.J. Nersessian, and P. Thagard, eds., pp. 165-179.

Kuhn, T.S., 1970, *The Structure of Scientific Revolutions* (1962), University of Chicago Press, Chicago, IL, second edition.

Kuhn, D., 1991, *The Skills of Argument*, Cambridge University Press, New York.

Kuhn, D., 1996, Is good thinking scientific thinking? in: D.R. Olson and N. Torrance, eds., pp. 261-281.

Kuipers, B.J., 1986, Qualitative simulation, *Artificial Intelligence* 29:280-338.

Kuipers, B.J., 1987, Qualitative simulation as causal explanation, *IEEE Transactions on Systems, Man, and Cybernetics* 17:432-444.

Kuipers, T.A.F., 1998, Abduction aiming at truth approximation, paper presented at the *International Conference on Discovery and Creativity* (ICDC), Ghent, Belgium, 1998.

Kulkarni, D. and Simon, H.A., 1988, The process of scientific discovery: the strategy of experimentation, *Cognitive Science* 12:139-176.

Kunstaetter, R., 1986, Intelligent physiologic modeling. Technical Report, Mit-Lcs-Tr-369, Laboratory for Computer Science, MIT, MA.

Lakatos, I., 1970, Falsification and the methodology of scientific research programs, in: *Criticism and the Growth of Knowledge*, I. Lakatos and A. Musgrave, eds., Cambridge University Press, Cambridge, pp. 91-195.

Lakatos, I., 1971, History of science and its rational reconstructions, in: *PSA 1970: in Memory of Rudolf Carnap*, R. Buck and R.S. Cohen, eds., Reidel, Dordrecht.

Lakatos, I., 1976, *Proofs and Refutations. The Logic of Mathematical Discovery*, Cambridge University Press, Cambridge.

Langley, P., Simon, H.A., Bradshaw, G.L., and Zytkow, J.M., 1987, *Scientific Discovery. Computational Explorations of the Creative Processes*, MIT Press, Cambridge, MA.

Lanzola, G., Stefanelli, M., Barosi, G., and Magnani, L., 1990, NEOANEMIA: a knowedge-based system emulating diagnostic reasoning, *Computer and Biomedical Research* 23:560-582.

Leake, D.B., 1992, *Evaluating Explanations. A Content Theory*, Erlbaum, Hillsdale, NJ.

Lenat, D., 1982, Discovery in mathematics as heuristic search, in: *Knowledge-Based Systems in Artificial Intelligence*, R. Davis and D.B. Lenat, eds., McGraw Hill, New York, NY.

Le Roy, E., 1899, Science et philosophie, *Revue de Méthaphysique et de Morale*, VII.

Levesque, H.J., 1989, A knowledge-level account of abduction, in: *Proceedings of the Eleventh IJCAI*, Morgan Kaufman, Los Altos, CA., pp. 1061-1067.

Levi, I., 1991, *The Fixation of Belief and Its Undoing*, Cambridge University Press, Cambridge.

Levi, I., 1996, *For the Sake of the Argument. Ramsey Test Conditionals, Inductive Inference, and Nonmonotonic Reasoning*, Cambridge University Press, Cambridge.

Leyton, M., 1989, Inferring causal history from shape, *Cognitive Science* 13:357-387.

Leyton, M., 1992, *Symmetry, Causality, Mind*, MIT Press, Cambridge, MA.

Lifschitz, V., 1986, Pointwise circumscription, in: *Proceedings of AAAI86*.

Lindsay, R.K., 1994, Understanding diagrammatic demonstrations, in: *Proceedings of the 16th Annual Conference of the Cognitive Science Society*, A. Ram and K. Eiselt, eds., Erlbaum, Hillsdale, NJ, pp. 572-576.

Lindsay, R.K, 1998, Using diagrams to understand geometry, *Computational Intelligence*, 9(4):343-345.

Lindsay, R.K., Buchanan, B., Feingenbaum, E., and Lederberg, J., 1980, *Applications of Artificial Intelligence for Organic Chemistry: The Dendral Project*, McGraw Hill, New York, NY.

Lipton, P., 1991, *Inference to the Best Explanation*, Routledge and Kegan Paul, London.

Lloyd, J.W., 1987, *Foundations of Logic Programming*, second edition, Springer, Berlin.

Lobachevsky, N.J., 1829-1830, 1835-1838, *Zwei geometrische Abhandlungen, aus dem Russischen bersetzt, mit Anmerkungen und mit einer Biographie des Verfassers von Friedrich Engel*, B.G. Teubner, Leipzig. Originally published as O nachalakh geometrii, *Kasanki Vestnik*, Feb.-March, 1829: 178-187; April, 1829: 228-241; Nov.-Dec., 1829: 227-243; March-April, 1830: 251-283; July-Aug., 1830: 571-636, and Novye nachala geometrii, *Uchonia sapiski Kasanskava Universiteta* 3, 1835: 3-48; 2, 1836: 3-98; 3, 1836: 3-50; 1, 1837: 3-97; 1, 1838: 3-124; 3, 1838: 3-65.

Lobachevsky, N.J., [1835] 1897, The introduction to Lobachevsky's New Elements of Geometry, translated from the Russian, with a preface by G.B. Halsted, *Transactions of Texas Academy* 2:1-17. Originally published in N.J. Lobachevsky, Novye nachala geometrii, *Uchonia sapiski Kasanskava Universiteta 3*, 1835: 3-48.

Lukasiewicz, J., 1970, Creative elements in science [1912], in: J. Lukasiewicz, *Selected Works*, North Holland, Amsterdam, pp. 12-44.

Lukaszewicz, W., 1990, *Non-Monotonic Reasoning. Formalization of Commonsense Reasoning*, Chichester, Horwood.

Lycan, W., 1988, *Judgment and Justification*, Cambridge University Press, Cambridge.

Magnani, L., 1988, Epistémologie de l'invention scientifique, *Communication and Cognition* 21:273- 291.

Magnani, L., 1991, *Epistemologia applicata*, Marcos y Marcos, Milan.

Magnani, L., 1992, Abductive reasoning: philosophical and educational perspectives in medicine, in: *Advanced Models of Cognition for Medical Training and Practice*, D.A. Evans and V.L. Patel, eds., Springer, Berlin, pp. 21-41.

Magnani, L., 1997, Basic science reasoning and clinical reasoning intertwined: epistemological analysis and consequences for medical education, *Advances in Health Sciences Education* 2: 115-130.

Magnani, L., 1999a, Creative abduction and hypothesis withdrawal in science, in: *Methodological Aspects of Discovery and Creativity*, J. Meheus and T. Nickles, eds., Kluwer, Dordrecht, in press.

Magnani, L., 1999b, Visual abduction in mathematical discovery, Technical Report, Department of Philosophy, University of Pavia.

Magnani, L., 1999c, Model-based creative abduction, in: L. Magnani, N.J. Nersessian, and P. Thagard, eds., pp. 219-238.

Magnani, L., 1999d, Inconsistencies and creative abduction in science, in: S. Colton, ed., pp. 1-8.

Magnani, L., 2000, *Philosophy and Geometry. Theoretical and Historical Issues*, Kluwer Academic, Dordrecht, forthcoming.

Magnani, L., 1991, ed., *Conoscenza e matematica*, Marcos y Marcos, Milan.

Magnani, L., Civita, S., and Previde Massara, G., 1994, Visual cognition and cognitive modeling, in: *Human and Machine Vision: Analogies and Divergences*, V. Cantoni, ed., Plenum, New York, pp. 229-243.

Magnani, L. and Gennari, R., 1997, *Manuale di logica. Logica classica e del senso comune*, Guerini, Milan.

Magnani, L., N.J. Nersessian, and Thagard, P., eds., 1999, *Model-Based Reasoning in Scientific Discovery*, Kluwer Academic/Plenum Publishers, New York.

Marcus, S., ed., 1988, *Automating Knowledge Acquisition for Expert Systems*, Kluwer, Dordrecht.

Marcus, S., 1988a, SALT: a knowledge-acquisition tool for purpose and revise systems, in: S. Marcus, ed., pp. 81-123.

Marr, D., 1982, *Vision*, Freeman, San Francisco.

Marr, D. and Nishihara, H.K., 1978, Representation and recognition of the spatial organization of three-dimensional shapes, in: *Proceedings of the Royal Society of London*, pp. 269-294.

Massaro, D.W., 1987, *Speech Recognition by Ear and Eye: A Paradigm for Psychological Inquiry*, Erlbaum, Hillsdale, NJ.

Matthews, M.R., 1989, *The Scientific Background to Modern Philosophy*, Hackett, Indianapolis/Cambridge.

Mayo, D., 1996, *Error and the Growth of Experimental Knowledge*, The University of Chicago Press, Chicago and London.

McCarthy, J. and Hayes, P.J., 1969, *Some philosophical problems from the standpoint of artificial intelligence*, in: *Machine Intelligence 4*, B. Meltzer and D. Michie, eds., Elsevier, New York, pp. 463-502.

McDermott, D., 1982, A temporal logic for reasoning about present and plans, *Cognitive Science* 6:101-155.

McDermott, D., 1988, Preliminary steps toward a taxonomy of problem solving methods, in: S. Marcus, ed., pp. 225-256.

McGraw, G. and Hofstadter, D.R., 1993, Perception and creation of alphabetic style, in: *Artificial Intelligence and Creativity: Papers from the 1993 Spring Symposium*, AAAI Technical Report SS-93-01, AAAI Press, Stanford.

Meheus, J., 1999, Model-based reasoning in creative processes, in: L. Magnani, N.J. Nersessian, and P. Thagard, eds., pp. 199-217.

Michalski, R.S., 1993, Inferential theory of learning as a conceptual basis for multistrategy learning, *Machine Learning* 11:111-151.

Mill, J.S., 1843, *A System of Logic*. Reprinted in: *The Collected Works of John Stuart Mill*, J.M. Robson, ed., Routledge and Kegan Paul, London.

Miller, A.I., 1984, *Imagery in Scientific Thought*, MIT Press, Cambridge, MA.

Miller, A.I., 1989, Imagery and intuition in creative scientific thinking: Albert Einstein's invention of the special theory of relativity, in: *Creative People at Work. Twelve Cognitive Case Studies*, D.B. Wallace and H.E. Gruber, eds., Oxford University Press, Oxford.

Milne, R., 1987, Strategies for diagnosis, *IEEE Transactions on Systems, Man, and Cybernetics*, 17:333-339.

Minski, M., 1985, *The Society of Mind*, Simon and Schuster, New York.

Mooney, R., 2000, Integrating abduction and induction in machine learning, in: P. Flach and A. Kakas, eds., pp. 181-192.

More, T., 1981, Notes on the diagrams, logic and operations of array theory, in: *Structures and Operations in Engineering and Management Systems*, P. Bjorke and O. Franksen, eds., Tapir Pub, Norway.

Morgan, M.S. and Morrison, M., eds., 1999, *Models as Mediators. Perspectives on Natural and Social Science*, Cambridge University Press, Cambridge.

Narayanan, N.H. and B. Chandrasekaran, 1991, Reasoning visually about spatial interactions, in: *Proceedings of the 12th International Joint Conference on Artificial Intelligence*, Morgan Kaufmann, Mountain View, CA, pp. 360-365.

Narayanan, N.H. and Motoda, M.S.H., 1995, Diagram-based problem solving: the case of an impossible problem, *Proceedings of 17th Annual Conference of the Cognitive Science Society*, Erlbaum, Hillsdale, NJ, pp. 206-217.

Nersessian, N.J., 1984, *Faraday to Einstein: Constructing Meaning in Scientific Theories*, Nijhoff, Dordrecht.

Nersessian, N.J., 1988, Reasoning from imagery and analogy in scientific concept formation, in: *PSA 1988,* A. Fine and J. Leplin, eds., vol. I, Philosophy of Science Association, East Lansing, pp. 41-47.

Nersessian, N.J., 1995a, Opening the black box: cognitive science and history of science. Technical Report GIT-COGSCI 94/23, July. Cognitive Science Report Series, Georgia Institute of Technology, Atlanta, GA. Partially published in: *Osiris* 10:194-211.

Nersessian, N.J., 1995b, Should physicists preach what they practice? Constructive modeling in doing and learning physics, *Science and Education* 4:203-226.

Nersessian, N.J., 1998, Kuhn and the cognitive revolution, *Configurations* 6:87-120.

Nersessian, N.J., 1999a, Inconsistency, generic modeling, and conceptual change in science, in: *Inconsistency in Science*, J. Meheus, ed., Kluwer, Dordrecht, in press.

Nersessian, N.J., 1999b, *Model-based reasoning in conceptual change*, in: L. Magnani, N.J. Nersessian, and P. Thagard, eds., pp. 5-22.

Nersessian, N.J, Griffith, T.W., and Goel, A., 1997, Constructive modeling in scientific discovery, Technical Report, Georgia Institute of Technology, Atlanta, GA.

Newell, A., 1982, The knowledge level, *Artificial Intelligence*, 18:82-106.

Newell, A., 1990, *Unified Theories of Cognition*, Harvard University Press, Cambridge, MA.

Newell, A. and Simon, H., 1963, GPS, A program that simulates human thought, in: *Computers and Thought*, E.A. Feingenbaum and I. Feldman, eds., McGraw Hill, New York, pp. 279-293.

Newton, I., 1721, *Opticks: Or a Treatise of the Reflections, Refractions, Inflections, and Colours of Light*, reprinted from the fourth edition with a foreword by A. Einstein and an introduction by E.T. Whittaker, G. Bell, London, 1931.

Newton, I., 1934, *Mathematical Principles of Natural Philosophy*, vol. I, translated by A. Motte and F. Cajori, University of California Press, Berkeley, CA.

Newton-Smith, W., 1980, *The Structure of Time*, Routledge and Kegan Paul, London.

Niiniluoto, I., 1999, Abduction and geometrical analysis. Notes on Charles S. Peirce and Edgar Allan Poe, in: L. Magnani, N.J. Nersessian, and P. Thagard, eds., pp. 239-254.

Oatley, K., 1996, Inference in narrative and science, in: D.R. Olson and N. Torrance, eds., pp. 123-140.

Oatley, K. and Johnson-Laird, P.N., 1987, Towards a cognitive theory of emotions, *Cognition and Emotions* 1:29-50.

Ohlsson, S., 1984, Restructuring revisited: summary and critique of the Gestalt theory of problem solving, *Scandinavian Journal of Psychology* 25:65-78.

Ohlsson, S., 1984, Restructuring revisited: an information processing theory of restructuring and insight, *Scandinavian Journal of Psychology* 25:117-129.

Okada, T. and Simon, H.A., 1997, Collaborative discovery in a scientific domain, *Cognitive Science* 21:109-146.

Olson, D. and Torrance, N., eds, 1996, *Modes of Thought. Explorations in Culture and Cognition*, Cambridge University Press, Cambridge.

O'Rorke, P., 1994, Abduction and explanation-based learning: case studies in diverse domains, *Computational Intelligence* 66:311-344.

O'Rorke, P., Morris, S., and Schulemburg, D., 1990, Theory formation by abduction: a case study based on the chemical revolution, in: J. Shrager and P. Langley, eds., pp. 197-224.

O'Rorke, P. and Ortony, A., 1992, Abductive explanations of emotions, in: *Proceedings of the 14th Annual Conference of the Cognitive Science Society*, Erlbaum, Hillsdale, NJ.

O'Rorke, P. and Ortony, A., 1995, Explaining emotions, *Cognitive Science* 18:283-323.

Ourston, D. and Mooney, R.J., 1994, Theory refining combining analytical and empirical methods, *Artificial Intelligence* 66:311-344.

Paivio, A., 1975, Perceptual comparisons through the mind's eye, *Memory and Cognition* 3(6):635- 674.

Patel, V.L., Evans, D.A., and Groen, G.J., 1989a, Biomedical knowledge and clinical reasoning, in: *Cognitive Science in Medicine. Biomedical Modeling,* D.A. Evans and V.L. Patel, eds., MIT Press, Cambridge, MA, pp. 53-112.

Patel, V.L., Evans, D.A. and Groen, G.J., 1989b, Reconciling basic science and clinical reasoning, *Teaching and Learning in Medicine* 1(3):116-121.

Patel, V.L., Evans, D.A. and Kaufman, D.R., 1989, A cognitive framework for doctor-patient interaction, in: *Cognitive Science in Medicine. Biomedical Modeling,* D.A. Evans and V.L. Patel, eds., MIT Press, Cambridge, MA, pp. 257-312.

Patel, V.L., Evans, D.A., and Kaufman, D.R., eds., 1990, Reasoning strategies and use of biomedical knowledge by students, *Medical Education* 24:129-136.

Patel, V.L. and Groen, G.J., 1991, The general and specific nature of medical expertise: a critical look, in: *Towards a General Theory of Expertise: Prospects and Limits,* A. Ericsson and J. Smith, eds., Cambridge University Press, Cambridge, pp. 93-125.

Patel, V.L., Groen, G.J., and Norman, G.R., 1993, Reasoning and instruction in medical curricula, *Cognition and Instruction* 10(4).

Patil, R.S, 1991, *Causal understanding of patient illness for electrolyte and acid-base diagnosis,* Technical Report, MIT-CSL-TR-267, Computer Science Laboratory, Massachusetts Institute of Technology, Cambridge, MA.

Pearl, J., 1988, *Probabilistic Reasoning in Intelligent Systems*, Morgan Kaufmann, San Mateo, CA.

Peirce, C.S., 1955a, Abduction and induction, in: C.S. Peirce, *Philosophical Writings of Peirce*, J. Buchler, ed., Dover, New York, pp. 150-156.

Peirce, C.S., 1955b, The fixation of belief, in: C.S. Peirce, *Philosophical Writings of Peirce*, J. Buchler, ed., Dover, New York, pp. 5-22.

Peirce, C.S., 1955c, Perceptual judgments, in: C.S. Peirce, *Philosophical Writings of Peirce*, J. Buchler, ed., Dover, New York, pp. 302-305.

Peirce, C.S., 1931-1958 (*CP*), *Collected Papers*, 8 vols., C. Hartshorne and P Weiss (vols. I-IV), and A.W. Burks (vols. VII-VIII), ed., Harvard University Press, Cambridge, MA.

Peng, I. and Reggia, I.A., 1987a, A probabilistic causal model for diagnostic problem solving I: integrating symbolic causal inference with numeric probabilistic inference, *IEEE Transactions on Systems, Man, and Cybernetics* 17:146- 162.

Peng, I. and Reggia, I.A., 1987b, A probabilistic causal model for diagnostic problem solving II: diagnostic strategy, *IEEE Transactions on Systems, Man, and Cybernetics*, 17:395-406.

Pennington, N. and Hastie, R., 1988, Explanation-based decision making: effects of memory structure on judgment, *Journal of Experimental Psychology: Learning, Memory, and Cognition* 14(3):521-533.

Pennock, R.T., 1999, *Tower of Babel. The Evidence Against the New Creationism*, MIT Press, Cambridge, MA.

Pennock, R.T., 2000, Can Darwinian mechanisms make novel discoveries? Learning from discoveries made by evolving neural network, in: Special Issue *Model-Based Reasoning in Scientific Discovery: Learning and Discovery*, L. Magnani, N.J. Nersessian, and P. Thagard, eds., *Foundations of Science*, forthcoming.

Pericliev, V. and Valdés-Pérez, R.E., 1998, Automatic componential analysis of kinship semantics with a proposed structural solution to the problem of multiple models, *Anthropological Linguistics* 40(2):272-317.

Piaget, J., 1952, *The Origins of Intelligence in Children*, International University Press, New York.

Piaget, J., 1974, *Adaptation and Intelligence*, University of Chicago Press, Chicago, IL.

Plato, 1977, *Plato in Twelve Volumes*, vol. II, Laches, Protagoras, Meno, Euthydemus, with an English translation by W.R.M. Lamb, Harvard University Press, Harvard (first printed 1924).

Poincaré, H., 1902, *La science et l'hypothèse*, Flammarion, Paris, (English translation by W.J.G. [only initials indicated], 1905, *Science and Hypothesis,* with a preface by J. Larmor, The Walter Scott Publishing Co., New York. Reprint, Dover Publications, New York, 1952).

Poincaré, H., 1905, *La valeur de la science*, Flammarion, Paris, (English translation by G.B. Halsted, 1958, *The Value of Science*, Dover Publications, New York).

Polanyi, M., 1966, *The Tacit Dimension*, Doubleday and Co., Garden City, New York.

Polya, G., 1957, *How to Solve It*, Doubleday/Anchor, Garden City, NY.

Poole, D., 1988, A logical framework for default reasoning, *Artificial Intelligence* 36:27-47.

Poole, D., 1989, Explanation and prediction: an architecture for default and abductive reasoning, *Computational Intelligence* 5:97-110.

Poole, D., 1991, Representing diagnostic knowledge for probabilistic Horn abduction, in: *Proceedings IJCAI-91*, Sydney, NSW, pp. 1129-1135.

Poole, D., 2000, Learning, Bayesian probability, graphical models, and abduction, in: P. Flach and A. Kakas, eds., pp. 153-168.

Poole, D., Goebel, R., and Alieliunas R., 1987, Theorist: a logical reasoning system for defaults and diagnosis, in: *The Knowledge Frontier: Essays in the Representation of Knowledge*, N. Cercone and G. McCalla, eds., Springer, Berlin, 1987, pp. 331-352.

Poole, D. and Rowen, G.M., 1990, What is an optimal diagnosis?, in: *Proceedings of the 6th Conference on Uncertainty in AI*, pp. 46-53.

Pople, H.E., 1973, On the mechanization of abductive logic, in: *Proceedings of the International Joint Conference on Artificial Intelligence 8*, pp. 147-152.

Pople, H.E., 1977, The formation of composite hypotheses in diagnostic problem solving, in: *Proceedings of the 5th IJCAI*, Morgan Kaufman, Los Altos, CA, pp. 1030-1037.

Pople, H.E., 1982, Heuristic methods for imposing structure on ill-structured problems: the structuring of medical diagnostics, in: *Artificial Intelligence in Medicine*, West View Press, Boulder, CO, pp. 119-190.

Pople, H.E., 1985, Evolution of an expert system: from INTERNIST to CADUCEUS, in: *Artificial Intelligence in Medicine*, I. De Lotto and M. Stefanelli, eds., Elsevier Science Publisher, Amsterdam, pp. 179-208.

Popper, K.R., 1959, *The Logic of Scientific Discovery*, Hutchinson, London, New York.

Popper, K.R., 1963, *Conjectures and Refutations. The Growth of Scientific Knowledge*, Routledge and Kegan Paul, London.

Port, R.F. and van Gelder, T., eds., 1995, *Mind as Motion. Explorations in the Dynamics of Cognition*, MIT Press, Cambridge, MA.

Port, R.F., Cummins, F., and McAuley, D., 1995, Naive time, temporal patterns, and human audition, in: R.F. Port and T. van Gelder, eds., pp. 339-371.

Powers, W.T., 1973, *Behavior: the Control of Perception*, Aladine, Chicago, IL.

Prigogine, I., 1980, *From Being to Becoming. Time and Complexity in the Physical Sciences*, Freeman, San Francisco, CA.

Prior, A., 1955, Diodoran modalities, *Philosophical Quarterly* 5:205-213.

Psillos, S., 2000, Between conceptual richness and computational complexity, in: P. Flach and A. Kakas, eds., pp. 59-74.

Pylyshyn, Z.W., 1981, The imagery debate: analogue media versus tacit knowledge, *Psychological Review* 88:16-45.

Pylyshyn, Z.W., 1984, *Computation and Cognition: Toward a Foundation for Cognitive Science*, MIT Press, Cambridge, MA.

Quaglini, S., Stefanelli, M., Barosi, G., and Berzuini, A., 1988, ANEMIA: an expert consultation system, *Computer and Biomedical Research* 21:307-323.

Quaglini, S., Berzuini, C., Bellazzi, R., Stefanelli, M., Barosi, G., 1989, Therapy planning by combining AI and decision theoretic techniques, in: *Proceedings of the Second European Conference of Artificial Intelligence*, pp. 147-156.

Quine, W.V.O., 1951, Two dogmas of empiricism, *Philosophical Review* 40:113-127. Also in: Quine, W.V.O., *From a Logical Point of View*, Hutchinson, London, 1953, 1961 (second edition), pp. 20-46.

Quine, W.V.O., 1979, *Philosophy of Logic*, Prentice-Hall, Englewood Cliffs, NJ.

Raisis, V., 1999, Expansion and justification of models: the exemplary case of Galileo Galilei, in: L. Magnani, N.J. Nersessian, and P. Thagard, eds., pp. 149-164.

Rajamoney, S.A., 1993, The design of discrimination experiments, *Machine Learning* 12:185-203.

Ram, A., Wills, L., Domeshek, E., Nersessian, N.J., and Kolodner, J., 1995, Understanding the creative mind: a review of Margaret Boden's *Creative Mind, Artificial Intelligence* 79:111-128.

Ramoni, M., Magnani, L., and Stefanelli, M., 1989, Una teoria formale del ragionamento diagnostico, in: *Atti del Primo Congresso della Associazione Italiana per l'Intelligenza Artificiale,* Cenfor, Genova, pp. 267-273.

Ramoni, M., Stefanelli, M., Magnani, L., and Barosi, G., 1992, An epistemological framework for medical knowledge-based systems, *IEEE Transactions on Systems, Man, and Cybernetics* 22(6):1361-1375.

Rayo, M., Gimate-Welsh, A., and Pellegrino, P., eds., 1997, *5th International Congress of the International Association for Semiotic Studies: Semiotics Bridging Nature and Culture*, Editorial Solidaridad, México.

Reggia, J.A., Dana, S.N., and Pearl, Y.W., 1983, Expert systems based on set covering model, *International Journal on Man-Machine Studies*, 19:443-460.

Reggia, J.A., Nau, D.S., 1984, An abductive non-monotonic logic, in: *Proceedings of the Workshop on Non-Monotonic Reasoning*, pp. 385-385.

Reichenbach, H., 1938, *Experience and Prediction*, University of Chicago Press, Chicago, IL.

Reilly, F.E., 1970, *Charles Peirce's Theory of Scientific Method*, Fordham University Press, New York.

Reiter, R., 1987, A theory of diagnosis from first principles, *Artificial Intelligence* 32:57-95.

Reiter, R. and De Kleer, J., 1987, Foundations of assumption-based truth maintenance systems: preliminary report, in: *Proceedings AAAI-87*, Seattle, WA, pp. 183-188.

Roch, I., 1983, *The Logic of Perception*, MIT Press, Cambridge, MA.

Roesler, A., 1997, Perception and abduction (abstract), in: M. Rayo, A. Gimate-Welsh, and P. Pellegrino, eds., p. 226.

Roussel, P., 1975, *PROLOG: Manual de référence et d'utilisation*, Group d'intelligence artificielle, Université d'Aix-Marseille, Luminy.

Saccheri, G., 1920, *Euclides Vindicatus. Euclid Freed of Every Fleck* (1733), translated by G.B. Halsted, Open Court, Chicago. Originally published as *Euclides ab omni naevo vindicatus*, Ex Typographia Pauli Antonii Montani, Mediolani (Milan).

Sakama, C., Abductive generalization and specialization, in: P. Flach and A. Kakas, eds., pp. 253-266.

Salmon, W.C., 1990, *Four Decades of Scientific Explanation*, University of Minnesota Press, Minneapolis.

Saltzman, E.L., 1995, Dynamics and coordinate systems in skilled sensorimotor activity, in: R.F. Port and T. van Gelder, eds., pp. 149-173.

Schaffner, K.F., 1986, Exemplar reasoning about biological models and diseases: a relation between the philosophy of medicine and philosophy of science, *Journal of Medicine and Philosophy* 11:63-80.

Schank, R., 1982, *Dynamic Memory: A Theory of Learning in Computers and People*, Cambridge University Press, Cambridge.

Schank, R. and Abelson, R., 1977, *Scripts, Plans, Goals and Understanding*, Erlbaum, Hillsdale, NJ.

Schmidt, H.G. and Boshuizen, P.A, 1992, Encapsulation of biomedical knowledge, in: *Advanced Models of Cognition for Medical Training and Practice*, D.A. Evans and V.L. Patel, eds., Springer, Berlin, pp. 265-281.

Scott, P.D. and Markovitch, S., 1993, experience selection and problem choice in an exploratory learning system, *Machine Learning* 12: 49-67.

Shanahan, M., 1989, Prediction is deduction but explanation is abduction, in: *Proceedings IJCAI-89*, Detroit, MI, pp. 1140-1145.

Shanahan, M., 1990, Abductive solutions to temporal projection problems, in: *Working Notes. AAAI Spring Symposium Series: Symposium on Automated Abduction*, Stanford University.

Shanahan, M., 1995, Default reasoning about spatial occupancy, *Artificial Intelligence* 74:147-163.

Shelley, C., 1996, Visual abductive reasoning in archaeology, *Philosophy of Science* 63(2):278-301.

Shelley, C., 1999, Multiple analogies in archaeology, *Philosophy of Science* 66:579-605.

Shepard, R.N., 1988, The imagination of the scientist, in: *Imagination and Education*, K. Egan and D. Nadaner, eds., Teachers College Press, New York.

Shepard, R.N., 1990, *Mind Sights*, Freeman, New York.

Shepherdson, J.C., 1984, Negation as failure: a comparison of Clark's completed data base and Reiter's closed world assumption, *Journal of Logic Programming* 1(1):51-79.

Shepherdson, J.C., 1988, Negation in logic programming, in: *Foundations of Deductive Databases,* Minker, J., ed., Morgan Kaufman, Los Altos, CA, pp. 19-88.

Shen, W.M., 1993, Discovery as autonomous learning from the environment, *Machine Learning* 12:143-165.

Shin, S.-J., 1994, *The Logical Status of Diagrams*, Cambridge University Press, Cambridge.

Shoham, Y., 1985, Ten requirements for a theory of change, *New Generation Computing* 3(4):467-477.

Shoham, Y., 1988, *Reasoning about Change, Time and Causation from the Standpoint of Artificial Intelligence*, MIT Press, Cambridge, MA.

Shoham, Y., 1989, Efficient reasoning about rich temporal domains, in: *Philosophical Logic and Artificial Intelligence*, Thomason, R.H., ed., Kluwer, Dordrecht, pp. 191-222.

Shoham, Y. and McDermott, D., 1988, Problems in formal temporal reasoning, *Artificial Intelligence* 36:49-61.

Shortliffe, E.H., 1976, *Computer-Based Medical Consultations*: MYCIN, Elsevier, New York.

Shrager, J. and Langley, P., eds., 1990, *Computational Models of Scientific Discovery and Theory Formation*, Morgan Kaufmann, San Mateo, CA.

Shunn, C. and Klahr, D., 1995, A 4-space model fo scientific discovery, in: *AAAI Symposium Systematic Methods of Scientific Discovery*, Technical Report SS-95-03, AAAI Press, Menlo Park, CA., pp. 40-45.

Simina, M.D. and Kolodner, J.L., 1995, Opportunistic reasoning: a design perspective, *Proceedings of the 17th Annual Cognitive Science Conference.*

Simon, H.A., 1965, The logic of rational decision, *British Journal for the Philosophy of Science* 16:169-186. Reprinted in: H. Simon, 1977, pp. 137-153.

Simon, H.A., 1969, *The Sciences of Artificial*, MIT Press, Cambridge, MA

Simon, H.A., 1976, Discussion: the Meno paradox, *Philosophy of Science* 43:147-151. Reprinted in: H. Simon, 1977, pp. 338-341.

Simon, H.A., 1977, *Models of Discovery and Other Topics in the Methods of Science*, Reidel, Dordrecht.

Simon, H.A., 1985, Artificial-intelligence approaches to problem solving and clinical diagnosis, in: *Logic of Discovery and Diagnosis in Medicine*, K.F. Schaffner, ed., University of California Press, Berkeley and Los Angeles, CA, pp. 72-93.

Simon, H.A., Valdés-Pérez, R.E., and Sleeman, D.H., 1997, special issue on *Scientific discovery*, *Artificial Intelligence* 91(2).

Steels, L., 1984, Second generation expert systems, *Journal on Future Generation Computers* 1:213-237.

Steels, L., 1990, Components of expertise, *AI Magazine* 12:38-49.

Stefanelli, M., Lanzola, G., Barosi, G., and Magnani, L., 1988, Modelling of diagnostic reasoning, in: *Modeling and Control in Biomedical Systems*, C. Cobelli and L. Mariani, eds., 163-174, Pergamon, Oxford, pp. 163-174.

Stefanelli, M. and Ramoni, M., 1992, Epistemological constraints on medical knowledge--based systems, in: *Advanced Models of Cognition for Medical Training and Practice*, D.A. Evans and V.L. Patel, eds., Springer, Berlin, pp. 3-20.

Stein, A.S., 1995, Imagination and situated cognition, in: K.M. Ford, C. Glymour, and P.J. Hayes, eds., 1995, pp. 167-182.

Stephanou, H. and Sage, A., 1987, Perspectives in imperfect information processing, *IEEE Transactions on Systems, Man, and Cybernetics* 17:780-798.

Stock, O., ed., 1997, *Spatial and Temporal Reasoning*, Kluwer Academic, Dordrecht.

Suárez, M., 1999, Theories, models, and representations, in: L. Magnani, N.J. Nersessian, and P. Thagard, eds., pp. 75-83.

Swanson, D.R. and Smalheiser, N.R., 1997, An interactive system for finding complementary literatures: a stimulus to scientific discovery, *Artificial Intelligence* 91(2):183-203.

Thagard, P., 1987, The best explanation: criteria for theory choice, *Journal of Philosophy*, 75:76-92.

Thagard, P., 1988, *Computational Philosophy of Science*, MIT Press, Cambridge, MA.

Thagard, P., 1989, Explanatory coherence, *Behavioral and Brain Sciences*, 12(3):435-467.

Thagard, P., 1992, *Conceptual Revolutions*, Princeton University Press, Princeton, NJ.

Thagard, P., 1996, *Mind. Introduction to Cognitive Science*, MIT Press, Cambridge, MA.

Thagard, P., 1997a, Collaborative knowledge, *Noûs* 31:242-261.

Thagard, P., 1997b, Coherent and creative conceptual combinations, Technical Report, Department of Philosophy, Waterloo University, Waterloo, Ontario, Canada.

Thagard, P. and Croft, D., 1999, Scientific discovery and technological innovation: ulcers, dinosaur extinction, and the programming language Java, in: L. Magnani, N.J. Nersessian, P. Thagard, eds., pp. 125-137.

Thagard, P. and Hardy, S., 1992, Visual thinking and the development of Dalton's atomic theory, in: *Proceedings of the 9th Canadian Conference on Artificial Intelligence*, Vancouver, pp. 30-37.

Thagard, P. and Shelley, C.P., 1997, Abductive reasoning: logic, visual thinking, and coherence, in: *Logic and Scientific Method*, M.L. Dalla Chiara, K. Doets, D. Mundici, J. van Benthem, eds., Kluwer, Dordrecht, pp. 413-427.

Thagard, P. and Verbeurgt, K., 1988, Coherence as constraint satisfaction, *Cognitive Science* 22:1-24.

Thagard, P., Gochfeld, D., and Hardy, S., 1992, Visual analogical mapping, in: *Proceedings of the 14th Annual Conference of the Cognitive Science Society*, Erlbaum, Hillsdale, NJ, pp. 522-527.

Thagard, P., Holyoak, K.J., Nelson, G., and Gochfeld, D., 1990, Analog retrieval by constraint satisfaction, *Artificial Intelligence* 46:259-310.

Thelen, E., 1995, Time-scale dynamics and the development of an embodied cognition, in: R.F. Port and T. van Gelder, eds., pp. 69-100.

Thinus-Blanc, C., Save, E., and Poucet, B., eds., Animal spatial cognition and exploration, in: N. Foreman and R. Gillett, eds., vol. II, pp. 59-86.

Thom, R., 1975, *Stabilité structurelle et morphogénèse. Essai d'une théorie générale des modélès* (1972), Reading, Benjamin, MA (English translation by D.H. Fowler, 1975, *Structural Stability and Morphogenesis: An Outline of a General Theory of Models*, W. A. Benjamin, Reading, MA).

Thompson, C.A. and Mooney, R.J., 1994, Inductive learning for abductive diagnosis, in: *Proceedings of the Twelfth national Conference on Artificial Intelligence*, Seattle, WA, pp. 664-669.

Torasso, P. and Console, L., 1989, *Diagnostic Problem Solving*, Van Nostrand Reinhold, New York.

Torretti, R., 1978, *Philosophy of Geometry from Riemann to Poincaré,* Reidel, Dordrecht, pp. 87-156.

Toth, I., 1991, *Essere e non essere: il teorema induttivo di Saccheri e la sua rilevanza ontologica*, translation by A. Marini, in: L. Magnani, ed., pp. 87-156.

Truesdell, C., *An Idiot Fugitive Essay on Science: Criticism, Training, Circumstances*, Springer, Berlin, 1984.

Turner, S.P., 1989, Tacit knowledge and the problem of computer modelling cognitive processes in science, in: *The Cognitive Turn. Sociological and Psychological Perspectives on Science*, S. Fuller, M. De Mey, T. Shinn, and S. Woolgar, eds., 1989, pp. 83-94.

Turner, S.R., 1994, *The Creative Process: A Computer Model of Storytelling and Creativity*, Erlbaum, Hillsdale, NJ.

Tversky, B., 1995, Cognitive origins of graphic productions, in: *Understanding Images. Finding Meaning in Digital Imagery*, F.T. Marchese, ed., Springer, New York, pp. 29-53.

Tversky, B., 1997, Memory for pictures, environments, maps, and graphs, in: *Intersections in Basic and Applied Memory Research*, D. Payne and F. Conrad, eds., Erlbaum, Mahwah, NJ, pp. 257-277.

Tweney, R.D., 1989, Fields of enterprise: on Michael Faraday's thought, in: *Creative People at Work. Twelve Cognitive Case Studies*, D.B. Wallace and H.E. Gruber, eds., Oxford University Press, Oxford.

Tweney, R.D., 1990, Five questions for computationalists, in: J. Shrager and P. Langley, eds., pp. 471-484.

Tye, M., 1991, *The Imagery Debate*, MIT Press, Cambridge, MA.

Valdés-Pérez, R.E., 1999, Principles of human computer collaboration for knowledge discovery in science, *Artificial Intelligence* 107(2):335-346.

Vila, L., 1994, A survey on temporal reasoning in artificial intelligence, *AI Communications*, 7(1):4-28.

Wason, P.C., 1960, On the failure to eliminate hypotheses in a conceptual task, *Quarterly Journal of Experimental Psychology* 23:63-71.

Wilson, P.N., 1997, Use of virtual reality computing in spatial learning research, in: N. Foreman and R. Gillett, eds., vol. I, pp. 181-206.

Winsberg, E., 1999, The hierarchy of models in simulation, in: L. Magnani, N.J. Nersessian, and P. Thagard, eds., pp. 255-269.

Winston, P.H. (1980), Learning and reasoning by analogy, *Communications of ACM*, 23(12).

Wirth, U., 1997, Abductive inference in semiotics and philosophy of language: Peirce's and Davidson's account of interpretation (abstract), in: M. Rayo, A. Gimate-Welsh, and P. Pellegrino, eds., p. 232.

Yamamoto, A., 2000, Using abduction for induction based on bottom-set generalization, in: P. Flach and A. Kakas, eds., pp. 267-279.

Zeigarnik, A.V., Valdés-Pérez, R.E., Temkin, O.N., Bruk, L.G., and Shalgunov, S.I., 1997, Computer-aided mechanism elucidation of acetylene hydrocarboxylation to acrylic acid based on a novel union of empirical and formal methods, *Organometallics* 16(14):3114-3127.

Zytkow, J., ed., 1992, *Proceedings of the ML-92 Workshop on Machine Discovery (MD-92)*, National Institute for Aviation Research, The Wichita State University, KA.

Zytkow, J., 1997, Creating a discoverer: autonomous knowledge seeking agent, in: J. Zytkow, *Machine Discovery*, Kluwer, Dordrecht, pp. 253-283, reprinted from *Foundations of Science* 2:253-283, 1995/96.

Zytkow, J., 1999, Scientific modeling: a multilevel feedback process, in: L. Magnani, N.J. Nersessian, and P. Thagard, eds., pp. 311-235.

Zytkow, J. and Fischer, P., 1996, Incremental discovery of hidden structure: applications in theory of elementary particles, in: *Proceedings of AAAI-96, AAAI Press*, Stanford, pp. 150-156.

Author Index

Abe, A. 36n
Abelson, R. 140n
Ackermann, R. 65
Addis, T.R. 58n
Akins, K. 44
Alchourrón, C. 24, 30
Aliseda, A. 21n, 32n, 48n
Allen, J.F. 119
Allwein, G. 104
Ampère, A.M. 57
Anderson, A. 128
Anderson, D.R. 19n, 42
Anderson, J.R. 103
Appelt, D. 104
Aravindan, C. 32n
Ayim, M. 19n

Bacon, F. 28, 29, 64
Bailer-Jones, D.M. 45n
Barosi, G. 94
Barrows, H.S. 93
Barsalou, L.W. 103
Barwise, J. 104
Bayes, T. 28
Becher, V. 30, 32, 33
Belnap, N. 128
Beltrami, E. 169
Bennet, B. 104
Benthem, J., van 119
Bergson, H. 46n, 117
Bessant, B. 36n
Besso, M. 118n
Beth, E.W. 48
Bhatta, S.R. 47n
Biot, J.B. 56, 56n, 59
Black, M. 40
Blades, M. 69
Bloch, H. 69
Blois, M.S. 19
Boden, M. 50
Boneh, D. 51n
Böök, A. 69
Boshuizen, P.A. 93, 94, 95
Boutilier, C. 30, 32, 33
Bovet, J. 69
Brewka, G. 33
Bringsjord, S. 48
Bradie, M. 11
Brown, J.R. 104
Bruner, J.S. 8, 43
Bruynooghe, M. 30
Buchanan, B.G. 28, 29
Bylander, T. 30

Carey, S. 143n
Carnap, R. 21n, 66
Chalmers, A.F. 65, 127
Chandrasekaran, B. 82-84, 105
Charniak, E. 16, 104
Chrisley, R.L. 68
Christiansen, H. 36n
Civita, S. 98
Clancey, W.J. 73, 82, 83, 91
Clark, A. 68
Clark, K.L. 147, 148, 148n
Colombetti, M. 68
Colton, S. 49n
Condillac, E. de 29
Console, L. 16, 30, 36, 36n, 37
Copernicus, P. 133
Cornuéjols, A. 46
Corruble, V. 51
Coz, P.T. 30
Crick, F.H. 98
Croce, B. 117
Croft, D. 129
Cross, C. 130n
Cummins, F. 121n
Curie, M. 159, 160

Dana, S.N. 16
Darden, L. 132, 133, 164, 166
Darwin, C. 41
Davies, J. 47n, 51
Davis, W.H. 19n, 42, 42n, 82
Davy, H. 55-57, 56n
Debrok, G. 19n
Degli Antoni, G. 113
de Kleer, J. 16, 30, 31, 33, 86
De Raedt, L. 30
Dimopoulos, Y. 37n
Dorigo, M. 68
Doyle, J. 24, 34
Dunbar, K. 50, 143, 144
Dung, P.M. 32n
Dunworth, C. 51n

Eco, U. 95, 140
Einstein, A. 39, 41, 98, 117, 118n, 124, 128, 135
Elden, A., van 65
Eshelman, L. 83, 84
Eshghi, K. 16
Etchemendy, J. 104
Euclid 91, 165-167
Evans, D.A. 73, 77, 80, 84, 89, 92, 93, 122
Evans, J. 77

Fajtlowicz, S. 50
Falkenhainer, B. 47n, 50, 104
Fann, K.T. 16, 19n
Faraday, M. 51, 56, 56n, 57, 59, 98, 163
Farah, M.Y. 97
Faye, J. 27, 41
Feltovich, P.J. 94
Feltzer, J.K. 17
Feyerabend, P. 27, 66, 118, 135
Feynman, R.P. 98
Finke, R.A. 97, 99, 141, 141n, 142, 142n
Fischer, P. 139
Fizeau, A.-H.-L. 135
Fizgerald, G.F. 162
Flach, J.M. 60
Flach, P. 18n, 21n, 36, 36n, 37
Fodor, J. 43n
Forbus, K. 47n, 51, 104
Foreman, N. 69
Fraassen, B., van 66
Frankfurt, H. 19n
Fresnel, A.J. 156
Freud, S. 126n, 149-152, 154, 155, 155n

Gadd, C.S. 73, 84
Galilei, G. 37n, 61, 63-65, 127, 133

Ganascia, J.-G. 51
Gärdenfors, P. 24, 30, 32, 33
Gärling, T. 69
Gauss, K.F. 168
Gennari, R. 24n
Gelder, T., van 68, 121n
Gentner, D.R. 47n, 104
Giaquinto, M. 104
Giedymin, J. 156, 158
Giere, R.N. 45n, 47
Gillet, R. 69
Ginsberg, M.L. 24n, 121n
Glasgow, I.J. 98-103, 111
Glymour, C. 7n, 50, 51n, 84
Gochfeld, D. 98, 104
Gödel, K. 49
Goel, A. 16, 47n, 51
Goldbach, C. 11, 12
Gooding, D.C. 16, 46, 54, 55, 56n, 58n
Goodnow, J.J. 8
Gorman, M.E. 19n, 52n
Graßoff, G. 51
Greenberg, M.J. 134, 165, 166
Gregory, R.L. 43n
Griffith, T. 51
Groen, G.J. 73, 79, 80, 89, 92-94, 95
Grünbaum, A. 149, 151, 154, 155

Habermas, J. 150, 151
Hacking, I. 65, 66
Hamilton, W.R. 156
Haneda, H. 36n
Hanks, S. 120
Hanson, N.R. 16, 17, 44, 65
Hardy, S. 98, 104, 106
Harman, G. 16, 17n
Harris, T. 45n
Hastie, R. 16
Hayes, P.J. 121
Hempel, C.G. 21, 39, 129
Hendricks, V.F. 27, 41
Hintikka, J. 2, 38, 39, 48, 129
Hinton, G. 99
Hobbs, J. 16, 104
Hoffrage, U. 28
Hofstadter, D.R 103
Holton, G. 98
Holyoak, K.J. 46, 47, 47n, 104
Hooke, R. 28, 29, 65
Hookway, C. 19n
Hunt, E. 79
Hutchins, E. 53, 59n, 61

Inoue, K. 36n
Ioerger, T.R. 105
Ippolito, M.F. 45n
Ironi, L. 85

Jackson, P. 25, 30
Jacobs, R.A. 105
Jenkins, M.A. 102
Johnson, L. 84
Johnson-Laird, P.N. 102, 103, 115n, 140, 145
Josephson, J.R. 16, 27, 30, 37, 38, 43n, 45n, 78, 104

Kakas, A. 16, 18n, 21n, 30, 36, 36n, 37, 37n
Kant, I. 9, 10, 10n, 42, 42n, 43n, 157
Kapitan, J. 19n
Katsuno, H. 130n
Kaufmann, W.J. 73, 80, 89
Kautz, H.A. 122
Kekulé, F.A. 98
Kepler, J. 19, 19n, 49
Keravnou, E.T. 84
Kirlik, A. 59, 60, 61n
Kirsh, D. 60
Klahr, D. 50, 143

Klein, F. 169
Koenig, O. 98, 165
Koestler, A. 98, 142n
Kolodner, J. 47n, 59n
Konolige, K. 30
Koslowski, B. 143
Kosslyn, S.M. 97-100, 102, 105, 165
Kowalski, R.A. 16, 30
Krauβ, S. 28
Kuhn, T. 27, 53, 65, 89, 118, 127, 128, 143n
Kuipers, B.J. 16, 84, 85
Kulkarni, D. 50
Kunstaetter, R. 91

Lakatos, I. 65, 66, 126-128, 126n, 133, 137, 162
Langley, P. 16, 49n, 50
Lanzola, G. 18, 85, 94
Laudan, L. 29
Leake, D. 127
Lenat, D. 49
Leonardo da Vinci, 126
Le Roy, E. 117
LeSage, G.-L. 29
Levesque, H. 25, 30
Levi, I. 30n, 127n
Leyton, M 103
Libiere, C. 103
Lifschitz, V. 122
Lindsay, R.K. 49, 105
Lipton, R.J. 17n, 51n
Lloyd, J.W. 148
Lobachevsky, N.J. 134, 167-169
Lorentz, H.A. 135, 162
Lovejoy, A. 117
Lukasiewicz, J. 19
Lukaszewicz, W. 24n
Lycan, W. 21

Mackworth, A.K. 16
Magnani, L. 10n, 16, 19, 24n, 33, 38, 41, 43, 69n, 92, 94, 98, 108, 130n, 134, 150, 169
Makinson, P. 24
Marcus, S. 82, 83
Markovitch, S. 50
Marr, D. 99, 104
Martignon, L. 28
Martin, P. 104
Massaro, D. 104
Maxwell, J.C. 39, 51, 156, 163
May, M. 51
Mayer, J.R. von 160
Mayo, D. 67
McAuley, D. 121
McCarthy, J. 121
McDermott, D. 82, 104, 119, 120
McGraw, A.P. 103
Meheus, J. 49n, 137
Mendelzon, A. 130n
Michalsky, R.S. 37n
Michelson, A.A. 39, 66, 135, 162
Mill, J. S. 20, 28
Miller, A.I. 98
Milne, R. 84
Minkowski, H. 117
Minski, M. 55n
Mittal, S. 82
Mooney, R.J. 37n,
Morange, F. 69
More, T. 99
Morgan, M.S. 45n, 59n
Morley, E.W. 39, 135, 162
Morris, G. 50
Morrison, M. 45n, 59n
Motoda, S.R.H. 105

Narayanan, N.H. 105
Nau, D.S. 16

Nersessian, N.J. 45, 45n, 47, 51, 98, 137n, 163, 163n, 169
Neumann, F.E. 156
Newell, A. 2, 49
Newton, I. 9, 29, 37n, 39, 117, 128
Newton-Smith, W. 119
Niiniluoto, I. 2, 48
Nishihara, H.K. 99
Norman, G.R. 92

Oatley, K. 16, 45, 45n, 140
Ockham, W. of 26
Oersted, H. C. 56, 57
Ohlsson, S. 14
Okada, T. 62n
O'Rorke, P. 16, 30, 37n, 45n, 50
Ortega y Gasset, J. 117
Ortony, A. 16, 45n
Ourston, D. 37n

Paivio, A. 97
Papadias, D. 98-102, 111
Patel, V.L 73, 79, 80, 89, 92-94, 95
Patil, R.S. 90, 94
Pearl, J. 16, 30, 135
Peirce, C.S. 2, 16, 17, 19-21, 19n, 21n, 22n, 25, 26, 35, 36, 41-46, 42n, 44n, 48, 49, 73, 95, 104, 107, 125, 140
Peng, I. 16
Pennington, N. 16
Pennock, R.T. 51n
Pericliev, V. 50
Piaget, J. 61, 68
Pietrsykowski, T. 30
Pizzi, P. 113
Plato 1, 3-8, 10, 12, 90
Poincaré, H. 133n, 145, 156-162, 157n, 169
Polanyi, M. 1, 8-11, 13, 90
Polya, G. 8
Poole, D. 16, 27, 30, 31, 33, 36n, 107n
Pople, H.E. 16, 90, 94,
Popper, K. 21, 28, 46n, 65, 66, 118, 126, 161, 162
Port, R.F. 68, 121
Poucet, B. 69
Powers, W.T. 60
Previde Massara, G. 98
Prigogine, I. 115, 116, 118, 118n
Prior, A. 119
Proclus 166, 167
Psillos, S. 36n
Putnam, H. 66
Pylyshyn, Z.W. 99

Quaglini, S. 85, 86
Quine, W.V.O. 34, 126

Raisis, V. 45n
Rajamoney, S.A. 50
Ram, A. 143n
Ramoni, M. 16, 18, 19, 73, 77
Reggia, J.A. 16
Reichenbach, H. 46n
Reilly, F.E. 19n
Reiter, R. 16, 31, 33
Remes, U. 2, 48
Riguzzi, F. 37n
Roch, I. 43n
Röntgen, W.K. 127
Roesler, A. 19n
Roussel, P. 148
Rowen, G.M. 27
Rutherford, E. 135

Saccheri, G. 166, 167
Sage, A. 24
Saitta, L. 36, 37
Sakama, C. 36n

Salmon, W.C. 21
Saltzmann, E.L. 68
Save, E. 69
Schaffner, K.F. 88
Schank, R. 140n
Schmidt, H.G. 93, 94, 95
Schulenburg, D. 50
Schunn, C. 50
Scott, P.D. 50
Sebeok, T.A. 95, 140
Selart, M. 69
Sgall, J. 51n
Shanahan, M. 30, 105, 122, 123
Shaw, R. 49
Shelley, C.P. 38, 43, 47n, 103, 107, 108, 130, 134, 138
Shen, W.M. 50
Shepard, R.N. 98
Shin, S.-J. 105
Shoham, Y. 116, 119-122
Shortliffe, E.H. 72n
Shrager, J. 16, 49n
Simina, M.D. 59n
Simon, H.A. 2, 7, 11-13, 16, 17, 24, 49n, 50, 62n, 81, 84, 90, 91, 95
Slayton, K. 97
Sleeman, D.H. 49n
Smalheiser, N.R. 50
Smith, S.M. 141n, 142, 142n
Socrates 3-5, 7-10, 12-14, 90, 91
Steels, L. 82
Stefanelli, M. 18, 73, 77, 84, 94
Stephanou, H. 24
Stein, L.A. 105n
Stickel, M. 104
Stock, O. 105
Suarez, M. 45n
Swanson, D.R. 50

Thagard, P. 16, 17n, 25-27, 29, 38, 40, 41, 43, 46, 47, 47n, 49, 50, 55n, 67n, 78, 98, 103, 104, 106, 126, 127n, 129, 129n, 130, 134, 136, 138, 144
Thelen, E. 68
Theseider Dupré, D. 30, 36n
Thinus-Blanc, C. 69
Thom, R. 119
Thomason, R.H. 130
Thompson, C.A. 37n, 40
Toni, F. 16, 30
Torasso, P. 16, 36n
Torretti, R. 167
Toth, I. 147n
Truesdell, C. 24
Turner, S.P. 3, 12, 14, 90
Turner, S.R. 139
Tversky, B. 105
Tye, M. 97, 99
Tweney, R.D. 45n, 49, 98

Valdéz-Pérez, R.E. 49, 49n, 50
Verbeurgt, K. 138
Vila, L. 122

Ward, T.B. 141n, 142, 142n
Warren, J.R. 60
Wason, P.C. 143
Wazon, J. 98
Whitehead, A. 117
Wilson, P.N. 113
Winsberg, E. 45n
Winston, P.H. 47n
Wirth, U. 19n
Witten, E. 138

Yamamoto, A. 36n

Zeigarning, A.V. 50
Zytkow, J.M. 45, 49n, 132, 139

Subject Index

abducibles, 36
abduction, 1, 2n, 15, 16, 17, 17n, 19, 20, 20n, 21, 22n, 23, 25-27, 29, 30, 33-39, 35n, 36n, 41-51, 42n, 45n, 53-57, 55n, 59, 60, 62n, 63, 65-69, 71-75, 78-81, 85-88, 91, 92, 94, 95, 97, 98, 104, 106-108, 110, 116, 123-126, 129-131, 133-135, 137, 139, 140, 143-145, 150, 161-165; and action, 55; action-based, 54; ampliative, 35; analogical, 49; and analogy, 45, 163; and analysis, 2n; and anomaly, 35, 125, 126; as answering, 129; and artificial intelligence, 16; and belief revision, 25; and causes, 37; cognitive model of, 71; and collaborative discovery, 62n; computational model of, 71; and conceptual change, 47, 133; as conjecture, 41; creative, 19, 25, 36, 39, 41, 42, 44, 46, 51, 57, 60, 67, 108, 129, 135, 161; and deduction, 21, 48; and disconfirming evidence, 143; and distributed reasoning, 144; double, 150; and emotion, 45, 45n, 140; and epistemic mediators, 59; and evaluation, 72-75, 78, 81, 85, 87, 88; and evaluation of hypotheses, 26, 27; existential, 49; and explanation, 36, 37, 39, 126; and formal (deductive) models, 30, 130; and generation of hypotheses, 1, 26; and geometrical proofs, 47, 48; and heuristics, 48; and holistic aspects of reasoning, 137; and hypothesis, 43; iconic, 43; and image-based explanation, 106; and inconsistency, 125, 126; and induction, 20, 36n, 57; and inductive generalization, 38; as inference, 35, 41, 42, 125; as inference to the best explanation, 25, 35, 73, 75, 78-81, 85-88, 92, 129; as inference to the best therapy, 75; and kinematic skills, 68; and language understanding, 104; and layered hypotheses, 134; logical account of, 130, 131; manipulative, 15, 16, 43, 51, 53, 55-57, 55n, 60, 63, 65-67, 140, 144; and medical education, 91; model-based, 15, 17n, 36, 38, 39, 41, 42, 42n, 44, 46, 51, 129, 133, 137, 163; and model-based heuristic, 48; and motor category formation, 68; and narratives, 34n, 139; and negation as failure, 145; of new facts, 65; and perception, 43, 104; and preferred explanations, 124; and preinventive forms, 140; and problem solving, 95; of psychoanalytic constructions, 150; rule-forming, 49; and science, 16; selective, 19, 23, 25, 29, 33, 57, 60, 72, 73, 75, 78-80, 85-87, 92, 108, 166;

and sense activity, 43; sentential, 29, 34, 35, 50, 130, 139; and sequential aspects of reasoning, 137; simple, 49; of spatial abilities, 69; and storytelling, 139; and surprising facts, 124; and the syllogistic framework, 17, 35; and templates, 54, 68; and template behavior, 66; temporal, 116, 123-124; and temporal explanation, 123; theoretical, 15, 17, 57, 59, 60, 67, 69; and thought experiment, 45; trans-paradigmatic, 41; and uncertanty, 79, 80; and unexpected findings, 143, 144; visual, 42, 46, 47, 97, 98, 104, 106-108, 110, 134, 163-168; and visual imagery, 46, 47, 98; and weak hypotheses, 145
abductive logic programming (ALP), 36
AbE, 50, 51
ABEL, 90, 94
abilities, 67-69; non-conceptual, 67; spatial, 67-69
AbMaL, 45
abnormality, 89
abstraction, 72, 73; of data, 72, 73
ACME/ARCS, 104
ACT-R, 50
action, 60, 61; and control of sense data, 61; epistemic, 60; and external artificial models, 61; and incomplete information, 61; and simplification, 61
adhocness, 161
AKO (a kind of), 100
AM, 49, 50
analogy, 21n
analysis, 2, 37n; and hypothesis generation, 2; and synthesis, 37n
ANEMIA, 85; architecture of, 85
anomaly, 17n, 30, 32, 33, 35, 36, 109, 124, 127 131-134, 136, 140, 143, 144, 162-164; conceptual, 133, 164-167; and conceptual change, 132; and conceptual problems, 134; confirmation of, 132; and disconfirming evidence, 140, 143; empirical, 131, 163; evaluation and assessment of, 132; and explanation, 127, 132; and failed predictions, 131; generation of, 132, 136; 136; and inconsistency resolution, 32; localization of, 132; and preinventive forms, 140; recognition of, 164; resolution of, 132, 133, 162-168; and theory change, 132; and unexpected findings, 140, 144
argument, 21, 22n; abductive, 21; as hypothesis, 22n; inductive, 21; nondefeasible, 21
arrays, 99-102, 105, 111-113; embedded, 101; hierarchy of nested, 111, 112; interpretation of, 102; multidimensional, 99, 102; nested, 100, 102; representation, 113; symbolic, 99
ARROSMITH, 50
artifact, 57, 61, 63, 64
artificial intelligence, 52, 77, 78; and cognitive psychology, 52; and epistemology, 52
ATMS, 86
atomic formula, 147n, 148

BACON, 50
Barbara, 22
basic science knowledge, 92, 94
Bayes' theorem, 28
belief, 21n, 25; degree of, 21n; logic of, 25
belief change, 39, 130; and conceptual change, 39, 130
belief contraction, 32
belief expansion, 32
belief revision, 24, 30, 32-34, 130; and inconsistency, 34
bifurcation phenomena, 119
bodies, 62; cognitive role of, 62; contact of, 167
branch-jumping, 40

CADUCEUS, 90, 94
CASNET, 90, 94

causal relationship, 20
causality of fate, 150
CDP, 51
CHARADE, 51
chronological ignorance, 121
circumscription, 123
clinical data, 149; probative value of, 149
clinical knowledge, 92, 94
clinical method, 154; as inductive, 154
clinical setting, 149; in psychoanalysis, 149
closed world assumption, 148n
cognition, 64, 97; distributed, 64; visual, 97
cognitive psychology, 52, 77, 78; and artificial intelligence, 52; and epistemology, 52
cognitive systems, 67, 68; in real time, 67; self-organizing, 68
coherence, 138; of conceptual systems, 138; and connectionist models, 138
coherence approach, 34
commodisme, 156
completed data base, 148, 148n
concept, 39, 40, 66, 87; and frames, 40; pathophysiological, 87; as tool, 66
concept learning, 18n
conceptual change, 39, 47, 130, 132, 135, 143n, 147, 162; and adding new concepts, 39, 130; and adding new instances, 39, 130; and adding new kind-relations, 39; and adding new part-relations, 39, 130; and adding new strong rules, 39, 130; and adding new weak rules, 39, 130; and anomaly, 132, 147; and belief change, 130; in childhood, 143n; and collapsing part of kind-hierarchy, 39; 130; and concept formation, 162; and incommensurability, 47; and ordinary thought, 143n; and reorganizing hierarchy by branch jumping and tree switching, 39, 130
conceptual combination, 128n, 129n; coherence-driven, 128n, 129n; and contradiction, 128n
conceptual problems, 134, 164; and anomaly, 134; as anomalies, 164; and formal sciences, 134
conceptual systems, 40
confirmation, 154; indirect, 154
confirmation theory, 21n
conflict, 126n
conjectural templates, 57
consilience, 26; and corroboration, 26
consistency, 33
constraint, 60; latent, 60
construal, 54, 55, 57, 59, 63
constructions, 3, 7, 9-11, 149-152, 154; confirmation of, 152; extension of, 152, 154; falsification of, 152; of figures, 3, 7; geometrical, 9-11; as histories, 150; most complete, 151; as narratives, 150; provisional, 151; in psychoanalysis, 152; withdrawing, 149, 151
contradiction, 34, 127, 128, 135, 145
convenient principle, 157, 157n
conventions, 146, 156, 157n, 160; as extensions of experimental laws, 160; geometrical, 157n; withdrawing, 156
conventional curriculum (CC), 92, 93
conventional principles, 156, 157n, 158-160; as convenient, 156; and experience, 156; and experimental laws, 158; extension of their experimental laws, 160; falsification of, 158, 159; unfalsifiability of, 158, 159
conventionalism, 133n, 138, 156, 161, 162; and falsificationalism, 133n, 161; generalized, 156; geometric, 156; nominalistic interpretation of, 161; of principles of physics, 156; and sophisticated falsificationism, 162
correctness, 27
counterinduction, 41, 135, 136
countertrasference, 150
creative process, 19
criteria, 26, 27, 33, 73, 88; epistemic, 27; ethical, 27; for evaluating hypotheses, 26;

evaluation, 73, 88; motivational, 27; of parsimony, 33; pragmatic, 27
CRUM (Computational-Representational Understanding of Mind), 67n

data base completion, 148n
deduction, 2, 18-24, 21n, 33, 37, 37n, 73, 75; and abduction, 21; and induction, 20, 37, 37n;
deductive reasoning, 17
defeasibility, 30
defeaters, 86
DENDRAL, 49, 50
dependency-directed backtracking (DDB), 34
depictionist view, 99
diagnosis, 19, 33, 73, 75, 77, 84, 85; consistency-based, 33; medical, 19
diagnostic reasoning, 8, 18, 21
diagrams, 63, 69, 166; cognitive, 69; and geometrical proofs, 166
diagrammatic reasoning, 42, 105
DIDO, 50
disconfirming evidence, 140, 143, 154; and anomaly, 143
discovery, 15, 16, 18; logic of, 15
discovery method, 126; coherence-driven, 126; data-driven, 126; explanation-driven, 126
disjunctive form, 147
dissipative structures, 118, 119
distributed resoning, 144
DNA computers, 51
dynamical systems, 67, 121

ECHO, 50
ecology, 59; cognitive, 59
economy, 27
electromagnetism, 56
embeddedness, 68
embodied thought, 68
embodiment, 54, 68
empirical adequacy, 27
encapsulation, 93
entrenchment, 32, 34
epistemic state, 30
epistemological model, 81, 84, 85
epistemology, 52, 77, 78; and artificial intelligence, 51; and cognitive psychology, 52
Euclidean postulate, 165-169; and experience, 165; fifth, 165-169
events, 69; unexpected, 69
event calculus, 123
evolution, 118
evolutionism, 117; metaphysical, 117
evolving neural networks, 51
exemplars, 71, 89; and generalization, 89
experiment, 63, 66, 67; and narratives, 55-57; scientific, 66
experiment design, 58n
experimental maps, 55
experimental narratives, 55-57
expert, 24
explanation, 17, 18, 20, 24, 27, 30, 73, 87, 106, 107, 119, 123, 124, 126, 127, 129, 132, 140; of action, 140; and anomalies, 126, 127, 132; and causes, 17; covering model of, 129; image-based, 106, 107; and inconsistency, 30; plausible, 20; preferred, 124; in temporal reasoning, 119
explanatory coherence, 27, 127n, 136
explanatory criteria, 27
explanatory hypothesis, 18
extended prediction problem, 121

facets, 73
fallacious reasoning, 17
fallacy of the affirming the consequent, 22
falsification, 128; and anomaly, 128
falsificationism, 133, 133n, 155, 161, 162; and anomaly resolution, 133; and conventionalism, 133n, 161; sofisticated, 133, 162; special, 155

findings, 140, 146; clinical, 146; unexpected, 140
foundation approach, 34
frames, 101
frame axiom, 121
frame problem, 121

GALATEA, 51
generalization, 22
generate and test, 2, 12
generic tasks, 82, 83
genetic algorithms, 51
geometrical proofs, 2-7, 48
geometry, 157, 158; Euclidean, 158; imaginary, 168; and physics, 157
Gestalt model, 35
Gestalt psychology, 8
gestures, 62n; cognitive role of, 62n
GLAUBER, 50
GRAFFITI, 50
GUIDON, 82, 83

heuristics, 2, 11, 29, 46, 48; and method of analysis and synthesis, 2; model-based, 48
heuristic classification, 82
heuristic pathway, 85
heuristic search, 28, 29, 49, 79
histories, 124; amended, 124
Horn clause, 147, 148
Horn clause theorem prover (query evaluation), 147, 148
hypotheses, 1, 2, 18, 18n, 19, 21-23, 21n, 25, 126, 128, 131, 133-135, 137, 139, 143, 145, 146, 160-162, 163; *ad hoc*, 133, 137, 143, 160, 161; and anomaly, 126; and anomaly resolution, 164; auxiliary, 137, 162; composite, 18; confirmation of, 21; diagnostic, 22, 23; and disconfirming evidence, 143; elementary, 18; evaluation of, 21n, 25, 26; and explanation, 135; explanatory, 18n, 25; and falsifiability, 162; formation and generation of, 1, 19, 21n, 25, 26, 143; and inconsistency, 126; inductive, 21n; justification, 25; layered, 134, 135; negation of, 146; new, 19; plausible, 19, 23, 26, 135; revision of, 143; as stratagems, 162; strong, 128, 139; and testability, 162; testing, 25, 143; unfalsifiable, 139, 145; weak, 139, 145; 131; withdrawing, 131, 145, 146
hypothesis generation, 1; and abduction, 1
hypothetico-deductive method, 21, 25

ID, 87
illness script, 93
imagery hypotheses, 107, 108
imagery objects, 110, 111; active, 111; passive, 111
imagery state, 113
imagination, 9, 10; in Kant, 9, 10
implementation, 84; computational method for, 84
incommensurability, 41, 47, 53, 55, 63, 67n; and conceptual change, 47
inconsistencies, 12, 34, 49n, 69, 124, 129, 130, 133, 135, 136, 138, 139; and connectionist models, 138; and curiosity, 129; epistemological, 136; and findings, 130, 133; generation of, 135; governing, 135; invention of, 136; logical, 136; maintaining, 136; and model-based reasoning, 49; and radical innovation, 135; resolution of, 32; story world, 139; and surprise, 129
inconsistency resolution, 32
induction, 18, 18n, 20-23, 21n, 25, 28, 29, 36-38, 36n, 37n, 74, 75, 86; and abduction, 20, 36, 36n; categorical, 22; confirmatory, 18n; and deduction, 20, 37, 37n, 38; descriptive, 18n; by elimination, 28, 29; and generalization, 20, 22, 36, 38; smart, 38
inductive generalization, 20n, 21n
inductive logic, 21n

inductive logic programming (ILP), 36
inductive reasoning, 17
inference, 4, 17, 17n, 19, 22, 27, 44, 44n; ampliative, 19; to the best explanation, 17, 17n, 19, 27; valid, 22
inferential commitment, 82, 94
influence diagrams (ID), 86
information, 61, 137; best possible, 137; incomplete, 61
instability, 124
interface, 61n; and reasoning, 61n; technological, 61n
interpretation, 149-151; and heuristics, 151; in psychoanalysis, 149-151
invariants, 139; formal, 139; observational, 139
irreversibility, 118

judgments, 157; synthetic *a priori*, 157

KBS (Knowledge-Based System), 71, 72, 82-85, 88, 90, 91, 94; medical, 71, 72, 82-85, 88, 90, 91, 94
KEKADA, 50
kind-hierarchy, 40, 41
kind-relation, 40

law covering model of scientific explanation, 39, 129
learning, 91; as self-programming, 91; problem-oriented, 91
LetterSpirit, 103
literal, 147, 148
LIVE, 50
logic, 29, 30; nonmonotonic, 30; modal, 30; programming, 29
Logic Theorist, 49
logical data base, 147

machine discovery programs, 49-51
MAGI, 104
manifestation, 22; patognomonic, 22
manipulation, 54; and cognition, 53
maps, 69, 106; cognitive, 69, 106
MDP/KINSHIP, 50
MDX-II, 84
MECHEM, 50
mediating structures, 59n
mediators, 59, 61, 62, 64-67; bodily, 67; cognitive, 61, 62; epistemic, 59, 61, 62, 64-67
medical knowledge, 19
medical diagnosis, 19
medical education, 72, 89-92, 95; problem-oriented, 91; science-centred, 91
medical science, 71; basic, 71; clinical, 71
memory, 100, 101; long-term, 100, 101
Meno paradox, 7, 8, 8n, 11, 12; Simon's solution of, 11, 12
mental image, 99
mental imagery, 59n, 99; spatial, 99; visual, 99
minimum mutilation, 34
MINSTREL, 139, 140
modal logic, 30
model-figure, 2
models, 14, 45n, 59, 60, 82, 83, 88, 102, 103; construction of, 45n; deep, 82, 83; external, 60; in history and philosophy of science, 45n; mediating role of, 45n, 59; mental, 14, 45n, 102, 103; overlapping, 88; in science, 45n; shallow, 82, 83; surface, 82
model-based abduction, 17n
model-based diagnosis, 31
model-based reasoning, 15
modus tollens, 28
MOLE, 84
monitoring, 19, 75-77, 84
motor category formation, 68

narratives, 7, 55-57, 62, 140, 150; of detection, 140; experimental, 55-57, 140; and psychoanalytic constructions, 150

negation, 145, 159; weak, 159
negation as failure, 139, 146-149, 152, 153, 155, 160; and completeness, 149; and constructions, 152, 153; and conventions, 160; and conventional principles, 160; in diagnostic reasoning, 147; and soundness, 149; and undecidability, 149; and validity, 148
NEOANEMIA, 71, 85, 87, 90, 94; and the epistemological model (ST-MODEL), 85
NEOCRIB, 84
NEOMYCIN, 82
neopositivism, 21
new experimentalism, 65, 67
NIAL, 103
non-Euclidean geometries, 134, 164-169
nonmonotonic logic, 30
nonmonotonicity, 150, 154, 155

objects, 62; cognitive role of external, 62
observations, 64; manipulation of, 64
ontological commitment, 82, 90, 94
ontological level, 84
ontologies, 72, 85, 91; causal, 85

paradigms, 127; and anomaly, 127
part-relation, 40, 100
pattern, 17, 68
perception, 8, 43, 43n, 44n, 45, 60; control of, 60
persistence problem, 121
phenomena, 57; and action, 57; anomalous, 57; and artificial apparatus, 57; dynamical, 57
PHINEAS, 50, 51
phisician, 79-81, 93; expert, 79, 80, 93; intermediate, 79-81; novice, 79, 80
physics, 157; and geometry, 157
pictorialist view, 99
plausibility, 19n, 20
prediction, 120
preference criteria, 24
preinventive forms, 140-142
principles, 33, 156, 158, 159, 165; as abstractions from experience, 165; conventional, 156, 158, 159; in geometry, 165; of parsimony, 33
probability, 21n, 28, 30; conditional, 21n
problem-based learning curriculum (PBL), 92, 93
problem solving, 1, 2n, 8, 9, 11, 13, 16, 18, 89, 90, 95; as reasoned process, 13
problem solving methods, 82, 83
processes, 20, 99; ampliative, 20; of visual imagery, 99
PROLOG, 148
psychoanalysis, 149
Pythagorean theorem, 3

QSIM, 85, 86
qualification problem, 120, 121
qualitative physics, 113
questioning, 129; theory of, 129

realism, 65, 66; scientific, 65, 66
reason maintenance system (RMS), 34
reasoning, 2, 13, 15, 18, 21n, 23, 24, 26, 28, 30, 33, 36, 37n, 38, 42, 45-47, 59, 61, 71-73, 79-81, 86, 89, 92, 95, 110, 115, 119-122, 124, 140, 144, 147, 163; and action, 115; analogical, 46, 47, 140; artificial system of, 26; backward, 79-81, 92, 95; basic science, 72, 89; and belief revision, 24; case-based, 47; causal, 37n, 38; and change, 115; clinical, 72, 89; deductive, 2, 30; defeasible, 13, 73; diagnostic, 18, 33, 36, 71; distributed, 144; forward, 79-81, 92, 95; human, 26; hypothetical, 2, 21n, 24; from imagery, 46; from incomplete information, 23; medical, 28, 36; model-based, 15, 42, 45; and negation as failure, 147; non-deductive, 21n; nonmonotonic, 23, 24, 73, 86, 124; opportunistic, 59; and prediction, 120; and

qualification problem, 120; simplification of, 61; smart, 38; spatial, 110, 163; temporal, 115, 116, 119, 121, 122, 124; therapeutic, 86; in thought experiment, 46; visual, 45
recollection, 2, 3, 5, 7; doctrine of, 2, 3, 5, 7
relations, 40, 101, 102, 105; part-whole, 40; kind, 40; spatial, 101, 102, 105; topological, 102
relevant entailment, 128
reminiscence, 1, 8, 10, 90; doctrine of, 1, 8, 10, 90
representations, 103, 105, 106, 137; and information, 105; model-based, 137; multiple, 103; spatial, 105, 106
retroduction, 16, 17, 19, 22n
retroductive process, 16
robotics, 68

schematism, 1, 9, 10
scientific creativity, 126; and inconsistency, 126
scientific discovery, 46, 46n, 59, 98, 118; and irrationality, 46n; as reasoned process, 46; and visual imagery, 98
scientific reasoning, 129; and inconsistency, 129
scientific theories, 127, 128, 138, 139; and anomalies, 127, 128; commensurable, 138; comparison of, 127; and contradiction, 127; incommensurable, 138; intertraslatable, 138; and invariants, 139; observationally equivalent, 138
search, 29, 153; unsuccessful, 153
segregated knowledge, 63
select and test, 91
semantic tableaux, 48n
sensation, 45
sense data, 61; control of, 61
sensory-motor period, 68
sentential model, 15
sentential framework, 38; limitation of, 38
simplicity, 26
simultaneity, 117
SME, 104
space, 117; perception of, 168; and time, 117
space-time, 117; four-dimensional, 117
spatial imagery, 162
spatial worlds, 110, 111, 113, 114; hierarchy of, 111; history of, 114; and navigation device, 110, 111; temporality of, 114
ST-MODEL, 17, 18, 20, 24, 28, 71, 72, 77, 81, 84, 85; and NEOANEMIA, 85
strategic principles, 48
strong-rule, 40
suggestion, 149; in psychoanalysis, 149
syllogism, 21, 21n, 105; and spatial reasoning, 105; statistical, 21n
syllogistic framework, 17, 20, 35
symbolic array theory, 99-101, 103
synthesis, 2, 37n, 42; and analysis, 37n; in Kant, 42
synthetic *a priori* judgments, 9, 10

TA, 86
tacit dimension, 8
tacit inference, 54
tacit knowledge, 1, 8-10, 90; and Kantian schematism, 10
task specific architecture, 82
teaching, 91; problem-oriented, 91
templates, 54, 57, 66, 68; conjectural, 57; non-conceptual, 68
temporal logic, 120
temporal reasoning, 121; and classical mechanics, 121
TETRAD, 50, 51
theory, 89; and domain of application, 89
theory change, 19, 28; in science, 19, 28
theory-ladenness, 16, 65
Therapy Advisor, 86, 87
therapy planning, 19, 73, 75-77, 84, 86
time, 116-121; and abduction, 116, 123-124; absolute, 117; and common sense rea-

soning, 120; direction of, 118; and dynamical systems, 121; and intuition, 118; and irrationality, 118; irreversible, 117; logic of, 119, 120; ontology of, 119, 120; and planning, 119; and prediction, 119; reversible, 116; and science, 116; and space, 117; universal, 117
tool, 66; as concept, 66
TORQUE, 51
training, 92; clinical, 92; situated, 92
transference, 149, 150; in psychoanalysis, 149
treatment, 75; medical, 75
tree-switching, 40
trial and error, 16, 81
true opinions, 3n
truth maintenance system (TMS), 34
tunnel effect, 46
tutoring systems, 91

uncertainty, 20, 24, 79, 87
unexpected finding, 144; and falsification, 144

VAMP.1, 104
VAMP.2, 104
visual buffer, 99
visual cognition, 97
visual imagery, 97-99, 102, 103, 106, 163; and problem solving, 106
visual thinking, 97, 98, 104, 134, 163; and analogy, 104, 166-168

weak-rule, 40
"world of paper", 63, 64

Yale shooting problem, 122, 123; bloodless variation of, 123