STRUCTURES IN SCIENCE

HEURISTIC PATTERNS BASED ON COGNITIVE STRUCTURES

An advanced textbook in neo-classical philosophy of science

by

Theo A.F. Kuipers

11 COMPUTATIONAL PHILOSOPHY OF SCIENCE[1]

Introduction

Computational philosophy of science is a collaboration between the philosophy of science and cognitive science. Cognitive science is itself a joint venture involving cognitive psychology, artificial intelligence research, neuroscience, linguistics, anthropology, and philosophy (Stillings et al., 1987), suggesting many interesting problems for philosophers of science (see Bechtel, 1988a). As illustrated in Chapter 6, Bechtel pays quite a lot of attention to reductionistic and non-reductionistic possibilities for co-operation between the disciplines involved.

In the computational philosophy of science, however, the roles have been interchanged. One tries to solve classical problems in the philosophy of science with means that have been particularly developed in cognitive psychology and artificial intelligence research. The kind of results aimed at are computer programs that enable certain cognitive tasks, or at least to simulate them, such as, discovering laws from data, designing hypotheses, evaluation and revision, concept formation, proposing experiments, etc.

The core ideas of the computational philosophy of science conceived as a collection of related, specific research programs are the following.

1) The most revolutionary starting point, viewed from classical philosophy of science, is that not only the evaluation process, the Context of Justification (CoJ), but also the discovery process, the Context of Discovery (CoD), is considered to be methodologically examinable and perhaps partially programmable. One even holds the classical distinction of contexts as meaningless since in the CoD one anticipates the CoJ, whereas the CoJ bears all kinds of traces from the CoD. However, as has generally been acknowledged, it is also here necessary to weaken some of the traditional CoJ-aims of the logical empiricists in a Popperian direction: certainty about knowledge claims cannot be obtained, for all judgements are provisional, and the hypothesis which is preferred at a certain moment need not be the only one compatible with the experimental data.

2) In the computational philosophy of science, scientific research is conceived as a form of problem solving. The co-operation between cognitive psychology and artificial intelligence has resulted in a general paradigm (developed by Newell and Simon) known as heuristic search (Newell and Simon, 1972). In a problem space of possible states, a problem is defined as the difference between an initial state and a goal state. State transitions are made by "if... then do ---"-rules, called production rules. In these rules heuristic considerations are built in, for which reason they are also called heuristic operations. A computer program for a certain type of problem hence essentially consists of a suitable specification of the problem space, in which the initial state (e.g., the data) and the goal state (e.g., a law describing the data) can be encoded, together with a set of heuristic operations.

3) There are at least four claims one may associate with such programs: historical adequacy, the claim to be able to reproduce milestones in the history of science; psychological

adequacy, the claim to be able to simulate in some relevant detail the processes in the minds/brains of scientists; philosophical adequacy, the claim to generate sound descriptive and explanatory theories that are increasingly successful, in empirical respects or even in approaching the truth; practical relevance, the claim to be of supportive value in the actual process of scientific discovery, evaluation, and revision. It is clear that these claims can be separately sustained, they do not need each other. However, as we will see, they may come into conflict with one another.

We will not pay attention to the implementation of programs, just a few remarks are to be made. Regarding program languages, it should be noted that the choice will not only depend on the specific goal, since the use of languages that can serve different purposes (such as LISP) has become of increasing importance, viz., in order to be able to connect different programs. Regarding techniques of knowledge representation, many possibilities for so-called expert systems have been developed. Davis (1990) provides a broad introduction to them.[2]

In Section 11.1. we will give an impression of a number of computer programs, among others, BACON, which searches for quantitative laws in physics, GLAUBER, STAHL and DALTON, which search for qualitative laws in chemistry, and (extensively) Paul Thagard's program PI, in which evaluation is at least as important as generation. Section 11.2. will critically deal with the evaluation part of Thagard's PI, his later program ECHO and his evaluative claims about it. It will be argued that implementation of the simple evaluation matrix, developed in Section 8.1., would lead to at least as good results as far as historical and philosophical adequacy of the resulting theory selections are concerned.

11.1. IMPRESSIONS ABOUT PROGRAMS

Introduction

We will start with the BACON-family of 'quantitative programs' (Subsection 11.1.1.1.), followed by three (mainly) qualitative programs (11.1.1.2.). Both collections of programs are extensively treated in (Langley et al., 1987). Next (Subsection 11.1.2.) we will present Thagard's PI (the abbreviation of Processes of Induction!) in some more detail (Thagard, 1988). In Subsection 11.1.3. we will give our impression about six examples of programs (presented in (Shrager and Langley, eds., 1990)) in which theory revision is central. Finally, Subsection 11.1.4. gives an indication of the possible interaction between the computational philosophy of science and certain kinds of philosophy and sociology of science.

The following survey specifies the nature and goals of the programs to be dealt with. It is important to note that theory formation may be data-driven, but may also be, in addition, theory-driven, and that theory revision usually starts with an anomaly, that is, a conflict between new data and the theory, whether or not in combination with the body of background knowledge, in order to be guided thereupon by (parts of) these three elements. Finally, all programs have something like separate theory evaluation, but not all deal with comparative evaluation.

	BACON	GLAUBER etc.	PI	Sect. 1.3
quantitative (mainly)	+	-	-	-
theory formation	+	+	+	-
theory revision	-	-	+	+
body of knowledge	-	-	+	+
comparative evaluation	-	-	+	+
further testing	-	-	-	+/-

Table 11.1.: Survey of nature and goal of programs

It is impossible to do justice in the following sketches to all the provisos that the authors (should) attach to their claims. For this reason, the exposition has a more optimistic tone than, strictly speaking, is justified. Nevertheless, we think that the programs illustrate a very interesting and important development.

11.1.1. The generation of quantitative and qualitative laws

11.1.1.1 In search of quantitative laws using the BACON-programs

The aim of the oldest programs, the BACON-programs, is the discovery of quantitative laws summarizing the available data, at least approximately. The starting point (initial state) is a set of real values of a number of independent and dependent variables. The goal is approached by means of the explicit definition of new terms on the basis of heuristic criteria, and successively testing whether or not a constant function or a linear relation has been obtained.

The basic heuristic operations are the following: if the data of two terms are inversely related, that is, if one increases and the other decreases, then the product of the terms is formed; if on the other hand the two terms are directly related, that is, the data increase or decrease together, the quotient is formed. It is important to note that these operations have been invented while designing the program, which illustrates the creative results that the computational philosophy of science may deliver.

In the following tables (derived from (Langley at al. 1987), p. 69, p. 85) the relevant steps are represented for designing Kepler's third law on the basis of fictitious or ideal data that satisfy the law (Table 11.2.1.) and on the basis of the original data (from Borelli) on the basis of which Newton checked whether the moons of Jupiter satisfied the law (Table 11.2.2.). In both cases it is easy to ascertain that the last term is equal to D^3/P^2 (D: the distance body-center, P: the period of the body's orbit), which Kepler's third law makes constant.

Planet	Distance D	Period P	Term-1=D/P	Term-2=DxT1	Term-3=T1xT2
A	1.0	1.0	1.0	1.0	1.0
B	4.0	8.0	0.5	2.0	1.0
C	9.0	27.0	0.333	3.0	1.0

Table 11.2.1.: Data obeying Kepler's third law of planetary motion

Moon	Distance D	Period P	Term-1=D/P	Term-2=DxT1	Term-3=T1xT2
A	5.67	1.769	3.203	18.153	58.15
B	8.67	3.571	2.427	21.035	51.06
C	14.00	7.155	1.957	27.395	53.61
D	24.67	16.689	.478	36.459	53.89

Table 11.2.2.: Borreli's data for Jupiter's satellites

The second table makes it immediately clear that a program cannot work with real data without rounding off criteria, as genuine empirical data never satisfy laws in a perfect way for various reasons. However, if the error percentage is too low BACON does not find a law and when it is too high 'false' laws are found. (Yet Kepler was satisfied with "D^2/P (Term-2) is constant" as a law for 10 years before he discovered "D^3/P^2 (Term-3) is constant"). Such problems arising from the need to approximate data do not get as much attention in (Langley et al., 1987) as might seem necessary.[3]

The BACON-programs, to which ever more heuristic operations have been added, have 'rediscovered' a large number of laws, sometimes on the basis of real data (indicated below by *) but more often on the basis of ideal data:

- starting from the indicated basic program: the laws of Kepler (*), Boyle (*), Ohm (*), and Galilei;
- starting from more than two variables and the addition of a stratified method of searching in the data, viz., varying different pairs of variables keeping the rest constant: the ideal gas law, and the laws of Coulomb, Kepler (refined), and Ohm (refined);
- starting from quantitative and nominal variables and an introduction method of intrinsic properties coupled to the nominal variable(s),
  - by means of a 'have a try method': the laws of Ohm (still more refined), with resistance as intrinsic concept, Archimedes (volume), Snellius (indices of refraction), Black (specific heat), conservation of momentum (inertial mass), and the law of gravitation (gravitational mass);
  - or by means of searching a common divisor: the results of Cannizaro (atomic weights) and Millikan (charge on the electron);

- assuming a theoretical point of view, hence (data- and) theory-driven, such as the search of symmetries or laws of conservation: refined versions of the laws of Snellius and Black, and the laws of conservation of momentum and energy.

11.1.1.2 In search of qualitative laws and models

It is plausible that the quantitative BACON-programs deal primarily with examples from physics. It is equally unsurprising that programs directed at qualitative laws and models focus in the first place on examples from chemistry. The three most well-known programs are even called after main figures in their history.

GLAUBER starts from a number of chemical reaction equations and properties (taste, and the like) of the participating substances, and forms classes of substances and reaction equations in terms of these classes of substances. In this way the classes now known as 'acids', 'alkalis' and 'salts' were formed as well as the law stating that "acids and alkalis produce salts", starting from the taste of eight substances and four reaction equations between them. The basic heuristic operations in the program amount to: form as large classes as possible, and quantify universally, if possible, and existentially otherwise.

STAHL also starts from a number of reaction equations, but is directed at the decomposition of compounds into their components. The heuristic operations are as simple as plausible: delete substances occurring at both sides of an equation; if at one side there is only one substance, then conclude that the other side specifies its components; and substitute a substance by its components in the remaining equations.

Finally, DALTON also starts from reaction equations (obeying Gay-Lussac's law of combining volumes, in spite of Dalton's personal rejection, see Chapter 1), and is directed at the determination of the (relative) number of molecules of each substance taking part in the reactions, and of the atomic formulas of the molecules of the participating substances. Here the heuristic operations are: the number of molecules have to be brought into agreement with Avogadro's hypothesis; the number of atoms at the left and the right side have to be equal; for the rest atomic formulas should be as simple as possible. It is clear that DALTON is strongly theory-driven and that new reaction equations may lead to a revision of atomic formulas. For example, from the single equation "hydrogen and oxygen generate water(vapor)" and the heuristic operations follow the molecular equation (using some plausible abbreviations) "2Hm(H-molecule) + 1Om $\rightarrow$ 2Wm" and the atomic formulas: 1Hm=1Ha(H-atom), 1Om=2Oa, and 1Wm=1Ha+1Oa; after adding the equation for the production of ammonia, viz., "hydrogen and nitrogen generate ammonia", DALTON revises its assumptions and finds the equations and formulas that we are used to.

The most plausible criticism of the programs presented so far is that the initial problem situation always presupposes a given conceptualization, a vocabulary, whereas the finding, or construction for that matter, of a suitable conceptualization usually is a very important part of the discovery process. However, the authors of the programs argue convincingly that the conceptualization that later turned out to be useful, was, as a matter of historical fact, in many cases available long before the discovery of the law was made. Hence, in these cases, the present type of (law-)discovery certainly has been of great historical importance.

It should finally be noted that the book by Langley et al. frequently makes nice excursions to so-called 'accidental discovery', and its programmability. Certainly with the techniques to be presented below, optimism in the latter respect is, on second thought, more realistic than at face value: notably, (parts of) programs for the detection of anomalies and the design of explanations by analogy can always be made sensitive to unsought findings.

11.1.2. Hypothesis formation and evaluation by PI

In this section we will first give an impression of the specifics of representation used in PI, followed by an indication of its various ways of induction. Next attention is given to the evaluation sub-program of PI. Finally, some ideas about Thagard's view of the philosophy of science from the computational point of view are given.

11.1.2.1. Representation

Whereas the type of programs dealt with in Subsection 11.1.1. belong to the mainstream in AI, Thagard's approach in his Computational Philosophy of Science (1988) is more related to cognitive psychology (CP). The point of departure is the distinctive representation of concepts, rules (for laws and hypotheses), theories and (explanatory) problems. A concept (e.g., sound) is characterized with the aid of super- and subordinate concepts (e.g., physical phenomenon, and music, respectively) and rules in which the concept occurs. In the following scheme (derived from Thagard, 1988, p. 17) the example of sound is represented:

Name: sound
Data-type: concept
Activation: 0
Superordinates: physical phenomenon, sensation
Subordinates: voice, music, whistle, animal sounds
Instances:
Activated by:

Rules:
    Rule-0: if *x* is heard then *x* is a sound
    Rule-1: if *x* is a sound then *x* is transmitted by air
    Rule-2: if *x* is a sound and *x* is obstructed, then *x* echoes
    Rule-3: if *x* is a sound and *y* is a person and *x* is near *y*, then *y* hears *x*
    Rule-4: if *x* is a sound then *x* spreads spherically
    Rule-5: if *x* is a sound then *x* is a sensation
    Rule-6: if *x* is a sound then *x* is physical phenomenon

Scheme 11.1.: A representation of the concept of sound

In Scheme 11.1. the rules are mentioned that are represented in extenso elsewhere in the systems represented in extenso. A rule postulates an empirical relation between two or more concepts (sound is propagated / reflected), represents a prototypical expectation, and is provided with a report of individual (explanatory) successes and counter-examples, and a degree of confidence based thereupon. A theory is a system of observation concepts (sound, wave), theoretical concepts (sound-wave), corresponding rules, general (explanatory) successes and counter-examples. Finally, there are two types of explanatory problems, viz., of an individual and general nature: the initial state concerns a specific or unspecific object (a certain sound or an arbitrary sound) and the goal state is supposed to give an explanation of behavioral features (propagation and reflection).

As all problems in 'CP-models' of heuristic search, explanatory problems are approached by PI in a quasi-connectionist way (see Subsection 11.2.1.2.), by activating the concepts that occur in the problem, followed by activation of concepts coupled to the first ones and then of the rules ('rule-firing'), for which the appropriate concepts are activated. This procedure is more or less standard in CP, and a problem is solved when some kind of match arises between the initial and the goal state. The special aspect of PI, as opposed to non-learning expert systems, is that, if the match does not occur, and only then, different kinds of induction, i.e., kinds of hypothesis and concept formation, become active, viz., generalization, abduction and concept formation. Such inductions are of course followed by evaluation.

11.1.2.2. Induction

Induction amounts to taking non-deductive steps. When *Ga* is to be explained by the initial state *Fa* and if no match occurs, the rule "if *Fx* then *Gx*" is formed by inductive generalization and subsequently evaluated on the basis of other information connected with *F* and *G*. When the rule is 'sufficiently and diversely supported', it is added to the 'knowledge base' with a corresponding degree of confidence (for more details, see Holland et al., 1986).

Abduction is a name covering four other forms of hypothesis formation, all followed by one and the same evaluation method, viz., a kind of so-called 'inference to the best explanation' to be discussed below. The four specific methods of abduction can be characterized as follows:

Individual abduction amounts to the proposal of the hypothesis *Fa* when *Ga* is to be explained and the rule "if *Fx* then *Gx*" belongs to the knowledge base and is activated, since *Ga* can be explained by this hypothesis and the rule mentioned. If it has to be explained that a certain sound is propagated and if the law "waves propagate" belongs to the base, the hypothesis is proposed that the particular sound is a wave.

Similarly, rule abduction amounts to the proposal of the hypothesis "if *Fx* then *Hx*" when the law "if *Fx* then *Gx*" has to be explained and "if *Hx* then *Gx*" belongs to the active knowledge base. In order to explain "sounds propagate" when "waves propagate" belongs to the base, the hypothesis "sounds are waves" is proposed.

It is clear that rule abduction is more or less a general form of individual abduction. Both would be examples of the fallacy known as 'affirming the consequent' if it were not the case that the hypotheses are merely proposed; they still have to be evaluated. The third form of abduction belongs to the same category of quasi-fallacies, the fourth form is of a quite different nature.

When the rule "if *S(x,y)* then *Gx*" belongs to the active knowledge base, and *Ga* is to be explained, existential abduction suggests the hypothesis "there is a *y* such that *S(a,y)*". Well-known examples are the postulation of the planet Neptune and of the substance phlogiston, and later oxygen. This form of abduction also occurs in metaphysics, think e.g., of postulating a human independent world.

The most surprising, and probably the least supported, hypotheses are formed by analogical abduction (AA): suppose that an individual or general fact *F** is to be explained and further that *F** is similar to *F* (according to certain activation patterns) and *F* is according to the knowledge base explained by *G*, then reconstruct the activation pattern around *F* and *G* is *-terms as much as possible, and postulate the corresponding analogue *G** of *G* as hypothesis.

Analogical abduction gives rise to a number of qualifying remarks. 1) If *G** passes the evaluation with success, PI subsequently forms an abstract scheme, with the pairs <*F,G*> and <*F*,G**> as instances, which makes a generalized form of analogical abduction (GAA) possible in the future. 2) There is also a specific causal form of AA (CAA): if *F* and *F** are similar, and *F* and *G* are causally related, and *G* and *G** are also similar, then the hypothesis is formed that *F** and *G** are causally related. Of course, here too a generalized form is possible. 3) Besides this special form of causal heuristics, science in general, and hence also PI, is guided by the general causal heuristics that a statistical correlation suggests a causal (cor)relation. This kind of heuristics is rather different from the causal heuristics that, according to Thagard, dominates in many pseudosciences: a similarity suggests a causal relation between those things that are similar. Note that the role of similarity is here of a quite different nature from that occurring in causal analogical abduction.

This last point completes the sketch of the four forms of abduction, which was, after inductive generalization, the second form of induction distinguished. The third and last form of induction concerns the formation of a theoretical concept and a corresponding theory. Suppose that in the context of a general explanatory problem an initially activated concept and a secondary activated concept share some instances, such that the rules coupled to these concepts are partly incompatible. In that situation one can form the new 'combined concept' and corresponding theory about the initial concept with the following expectations: insofar as all rules of both are compatible they are included, and of two incompatible rules the one belonging to the initial concept is included.

Recall that in the case of rule abduction the example was mentioned that the concept of sound becomes related to that of wave (of fluids), which leads to incompatible expectations, viz., spreading spherically versus in a single plane. In this case the concept of sound-wave is formed, with spherical propagation attached to it, and the wave theory of sound is formed, i.e., "sounds are sound-waves", with explanatory successes, among others, that sounds propagate and reflect.

11.1.2.3. Evaluation

PI will in general propose, if possible given the active knowledge base, several hypotheses and theories for the solution of an explanatory problem. These should be evaluated, where the knowledge base itself is also in discussion. More specifically, the evaluation by PI consists of a kind of 'inference to the best explanation' (IBE), in which not only the original problem takes part, but also the explanations that already belong to the base. Thagard deals with the theories of Darwin, Lavoisier and Fresnel in order to make clear that IBE has played an important role in the history of science.

Suppose that for the explanation of a new general fact at least one new hypothesis is formed. PI then determines all competing hypotheses (new or already in the base) and applies two criteria: consilience and simplicity. Thagard also argues that a third criterion, a kind of analogy score, should also be taken into account in an extended version of PI, but we have doubts about this, heuristics exceeding, role of analogy.

Two hypotheses are competing when they are logically independent and have overlapping success or when they are the ends of a chain of hypotheses with overlapping successes. All successes of all competing hypotheses are to be taken into account in the final evaluation.

A hypothesis is more consilient in proportion to the number (and the importance) of its successes. A hypothesis is the more simple the fewer auxiliary hypotheses need to be assumed to achieve its successes. If both criteria lead to the same ordering of the hypotheses, the most successful and the most simple hypothesis is chosen. If the orderings do not coincide, which is of course very well possible, the value of a hypothesis is defined as the product of its successfulness and its simplicity, and IBE now tells us that the hypothesis with the highest value is to be chosen.

Note that Thagard's notion of consilience is very similar to the notion of successfulness that we used in Chapter 7 and 8. Moreover, it is important to stress that Thagard uses 'simplicity' not in the usual sense as a matter of the structure of a hypothesis, but operationalizes it, very convincingly, in terms of the number of specific extra assumptions that have to be made for the successes. It is in this respect that Thagard also notes an important difference in his 'portraits' of science and pseudoscience.

IBE raises of course the question of its justification. Thagard spends two chapters on this problem. First he derives three criteria for the evaluation of systems of argumentation (consisting of normative principles, practices, background theories, and inferential goals) from methods of evaluation proposed for the evaluation of other systems of argumentation, among others, the method of 'reflective equilibrium' in Rawls' social philosophy. Thagard (1988, p. 129) arrives at the following three criteria for the coherence of an inferential or inductive system:

1. Robustness: to what extent do the normative principles account for inductive practice?
2. Accommodation: to what extent do available background theories account for deviations of inductive practice from the normative principles?
3. Efficacy: given background theories, to what extent does following the normative principles promote the satisfaction of the inferential goals?

According to Thagard, IBE is robust because it is used very frequently, as illustrated by the above mentioned examples. Deviations of IBE can be explained (accommodated) with the aid of psychological theories of motivation. And IBE is effective with respect to three goals: explanation, prediction and truth approximation. The first is evident, and the second is implied by the fact that prediction is a kind of potential explanation. For the third goal Thagard takes another detour by first arguing in favor of realism, at least for the natural sciences. He does so by applying IBE (in a non-circular way, or so he claims) on the relevant philosophical theories with respect to the explanation of three phenomena: technological applications, the accumulation of knowledge, and consensus about claims to knowledge. In view of our analysis of the relation between successfulness and truth approximation, the last detour, of which the non-circular character is not beyond doubt, is superfluous. More specifically, we have shown that IBE, provided it is revised in the form 'inference to the best theory as the closest to the truth', is demonstrably functional for truth approximation (ICR, Section 8.5.).

11.1.2.4. Varia

Thagard's book is rich in ideas about the philosophy of science looked at from the computational point of view. Starting from a non-extreme form of theory ladenness of observation terms, Thagard arrives at a moderate kind of meaning holism and a corresponding methodological conservatism, leaving room for the practical commensurability that has been built into PI, as we have seen, for the evaluation of hypotheses.

So-called evolutionary epistemology is according to Thagard based on a bad analogy. Regarding the three main problems (the explanation of variation, selection, and transmission) the analogy gives the wrong answers. Hence, PI may be supposed to assign a low value to the 'evolutionary hypothesis' concerning knowledge development if that would even be formed by analogical abduction at all.

The two differences already mentioned between science and pseudoscience form elements of two Wittgensteinian profiles that Thagard distinguishes, leaving room for intermediate forms: correlation versus resemblance thinking, being directed at empirical evaluation or not, comparison with alternative theories or not, aiming at as simple theories as possible or not, aiming at progress, and to begin with, aiming at the improvement of theories or not. (See also Section 8.3. for a related diagnosis of pseudoscience).

Thagard also argues that, applied to cognitive science, the computational approach might help to make a real science of it, with real successes. However, for that purpose it is necessary that PI, or similar programs dealing with problem solving, not be restricted to induction and evaluation of hypotheses, but that the design of experiments should also be taken into account. However, most of the (directed) connections between the three relevant activities, viz., problem solving, induction, experimenting, still have to be elaborated.

Thagard emphasizes the importance of parallel computational systems, because it leads to better programs. Moreover, Thagard suggests that we consider and simulate a scientific community as a collection of interacting parallel computational systems. Here the idea of social adequacy comes into the picture. The suggested analogy raises the question of the possibility and the desirability of a further division of labor than the well-known one between theoreticians and experimenters. In this connection he thinks in particular of the division of labor between 'constructive dogmatists' and 'skeptical critics', to use our favorite terms for what is essentially the same idea.

Finally, in the course of the book, all three forms of adequacy (historical, psychological, philosophical), distinguished in the introduction, play a comparably important role. In his later book, Conceptual Revolutions (1992), to be critically discussed in Section 11.2., historical adequacy is the main goal (an example, revolutions in geology, is contained in the collection of Shrager and Langley (1990), the collection to be dealt with in Subsection 11.1.3., and the example in Subsection 11.2.2.2.). Unfortunately, Thagard does not pay explicit attention to the fourth possible claim of computational philosophy of science, viz., practical relevance, although he regularly hints at the possibility of computer assisted discovery and evaluation.

11.1.3. Forms of theory revision

The first five chapters of the book edited by Shrager and Langley, Computational models of scientific discovery and theory formation (1990), deal with variations and elaborations of what has been presented so far (two of the chapters have already been indicated). The last five chapters deal with topics central to cognitive psychology. Our focus will be on chapters 6 to 11, which deal with important innovations in comparison to the foregoing, with emphasis on theory revision. The point of departure is always a knowledge base of data and theories, usually represented by the general method of representation of qualitative processes proposed by Forbus (Cf. Davis 1990). It should be noted in advance that there is not much similarity with the so-called 'belief revision' research program of Gärdenfors (1988) and others. The main reason seems to be that the latter program does not attempt to deal with scientific knowledge. In particular, it is 'actual' rather than 'nomic world' oriented, and it is consistency driven.

Falkenhainer has developed a program (PINEAS) that realizes a far-reaching unification between four scenarios of explanation (deductive, using assumptions, generalizing, or using an analogy) and corresponding theory formation and revision by viewing the latter as aiming at the maximally feasible similarity between the explanandum and the explanans. The further selection of theories takes place on the basis of the degree of severity of the extra assumptions (in increasing order): no additions, addition of new properties of an acknowledged type to already acknowledged entities, addition of new entities of an acknowledged kind, addition of new properties of a new type to acknowledged entities, and finally the addition of new entities of a new type.

The 'Abduction Engine' (AbE) devised by O'Rorke, Morris and Schulenburg pretends to make revolutionary revision of a theory possible, starting from an observational anomaly between the theory, the knowledge base, and new data. After the recognition of such an anomaly, the theory is weakened to a basic theory, which is compatible with the new data. Next, this basic theory is strengthened, by abduction, roughly along similar lines as in Thagard's PI, to a theory that is able to explain the anomalous data. To be sure, the weakening and subsequent strengthening corresponds to 'contraction' and 'expansion' in the belief revision program of Gärdenfors et al.. The AbE-program is developed on the basis of the transition of the phlogiston theory to the oxygen theory of combustion. Hence, it is no surprise that the program is able to reproduce that episode.

In the COAST program of Rajamoney theory revision is approached stepwise: 1) registration of an anomaly between new data and a theory in the knowledge base, 2) designing proposals for revision that transform the anomalous data into explanatory successes, 3) suggesting experiments, 4) first selection on the basis of the anomaly to be explained, the outcome of the experiments, and the degree in which previous successes are retained, 5) further selection on the basis of simplicity and predictive power. This program, and the following two, were guided in their development by examples from biochemistry.

The KEKADA program developed by Kulkarni and Simon registers surprising phenomena by comparing new experimental results obtained for 'normal' purposes with expectations on the basis of the knowledge base, and proceeds by dealing with such phenomena; that is, it tries to revise the relevant theory and proposes experiments to test such revisions. For this purpose five strategies are built into the program, each of which consists of a hypothesis generator, an experiment proposer and an evaluator. The core idea of, for example, the first strategy is to try to strengthen the surprising phenomenon by independently manipulating the variables of the system.

Karp's program (HYPGENE) considers hypothesis formation and revision as a matter of design, with constraints (a profile of desired properties) and operators that revise the provisional prototype (the profile of factual properties) in order to make it satisfy the constraints more and more. For a detailed localization of HYPGENE in the context of drug research, and its similarity to the models of design research presented in Chapter 10, see Bosch (2001, Section 9.5).

Finally, Darden considers the solution of an anomaly of a theory as a task for diagnostic reasoning, which has been developed for expert systems, guided by the tracing of a defect in a technical system. She shows, using Mendelian examples (see also Darden (1991)), that the decomposition of all presuppositions of a theory may provide the points of departure for solving the anomaly, and that its solution, as in Lakatos' famous analysis (1976), may or may not lead to fundamental theory revision, that is, a revision staying within a research program or breaking through the barriers of the relevant program.

11.1.4. Some connections with neo-classical philosophy of science

As far as modern philosophy of science is directed at cognitive heuristic patterns in knowledge and method, such as this book is, there are many conceivable connections with the computational philosophy of science. On the one hand these points of contact concern the methodological topics of separate and comparative evaluation and truth approximation, on the other hand they involve cognitive patterns that could provide the heuristics for computer programs.

In the computational literature about theory evaluation and revision, there remains quite a lot of confusion as to what exactly are the units of successes and problems, and how they hang together. Our analysis of theory evaluation (Chapter 7 and 8) provides several suggestions for what precisely is to be conceived as an explanatory success, what can be conceived as a descriptive problem of theories and test implications, and how they can be represented. The plausible 'instrumentalist' goal of theory revision, increase of explanatory success, and decrease of descriptive problems, is also made more precise in this way. In the above sections, formulations that were sharpened from this perspective have already been used from time to time. Moreover, in the paper "Abduction aiming at empirical progress or even truth approximation" (Kuipers, 1999b) we have elaborated the conceptual prospects for generalizing, on the basis of the analysis in ICR, the aim of abduction, showing how the standard goal of explaining surprising or anomalous observations can be integrated into a notion of empirical progress or even truth approximation. It turns out that the main abduction task then becomes the indicated instrumentalist task of theory revision with the intent of devising an empirically more successful theory, relative to the available data, but not necessarily compatible with them. The rest, that is, genuine empirical progress as well as observational, referential and theoretical truth approximation, is a matter of evaluation and selection, and possibly new generation tasks for further improvement. The paper concludes with a survey of possible points of departure, in AI (in particular, PI and the programs indicated in Subsection 11.1.3.) and logic, for computational treatment of the instrumentalist task guided by the 'comparative evaluation matrix'. Below we will briefly indicate one of the logical approaches, viz., the semantic tableau approach.

We hinted already at the relevance of our correction of 'inference to the best explanation' and the possibility of directly justifying that corrected version as functional for truth approximation.

Finally, it is indeed illuminating, as Karp suggests, to conceive theory formation and revision as a kind of design research, with the ultimate purpose of constructing the strongest true theory within a given or adapted vocabulary. As we have indicated in Chapter 10, the differences between descriptive and explanatory research programs on the one hand and 'normal' design research programs on the other are at least as interesting as the similarities.

The heuristic patterns that neo-classical philosophy of science can provide concern in the first place standardized decompositions of different kinds of explanation and reduction (Chapter 3-5). These can easily be reinterpreted as generalized abduction schemes in the sense of Thagard. Moreover, they can be used for diagnostic reasoning in case of anomalies as proposed by Darden.

Other, more global, heuristic patterns are provided by the analysis of the structure and development of research programs and their possibilities for interaction (Chapter 1), the hierarchy of knowledge arising from the theory relative distinction between observational and theoretical terms (Chapter 2), and the internal structure of theories (Chapter 2 and 12). For instance, it will become quite clear in the next chapter that the structuralist analysis of theories can almost directly be used for the computational representation of theories.

Let us turn to the promised brief discussion of one possible logical approach[4] to instrumentalist abduction, viz., the semantic tableau approach. It was introduced by Mayer and Pirri (1993) (see also (Pirri, 1995)), and redirected in a more efficient way and one more relevant to the philosophy of science by Atocha Aliseda (1997), who focusses on novelty and anomaly guided abduction. In the case of a surprising observation, i.e., the situation that new evidence *E* is not entailed by the background theory *B*, the task is to expand the latter with some hypothesis *H* such that *B&H* entails *E*, but *H* alone does not. This is called novelty guided abduction. In the case of an anomalous observation, i.e., the situation that *E* contradicts *B*, the task is to revise *B* to *B** such that *B** entails *E*. This is called anomaly guided abduction. We will only indicate the basic idea behind Aliseda's algorithms for dealing with these types of abduction. These algorithms are as yet restricted to the semantic tableaux for propositional languages. A semantic tableau is normally constructed (see e.g., Smullyan, 1968) to find out whether a set of premises logically entails a conclusion by systematically trying to construct a counter-example, that is, a propositional model or structure in which the premises are true and the conclusion is false. If all branches generated by the (formula decomposing) tableau rules from the set of premises and the negated conclusion become 'closed', that is, contain a formula and its negation, the argument is valid. Open branches describe counter-examples to the argument. In the case of novelty guided abduction, there must be open branches in the tableau starting from the background theory *B* and the negation of the phenomenon description *E*. Systematically closing these branches produces equally many consistent proposals for abductive explanation of the novel phenomenon. In the case of anomaly guided abduction, the background theory apparently entails the negation of the anomaly; that is, the corresponding tableau starting from the theory and the phenomenon description is closed. Now the task is to systematically open the branches of this tableau by deleting parts of the theory (contraction), followed by adding new formulas (expansion) that lead to the closing of all branches when combined with the negation of the phenomenon description.

The above description neglects all nuances and versions of Aliseda's approach, but it may nevertheless already be plausible to conjecture that the approach can be generalized to instrumentalist abduction purposes. The main thing is to operationalize in tableau terms the possibility that one theory is 'more incompatible' with the data than another. Then the task is to generate theories that are not more incompatible with the available data than the original one, explain at least as much as the latter, and improve the latter in at least one of these respects. Some other adaptations will be required to make it suitable for scientific purposes. First, room should be left so that general facts can be used as data and, second, the original and resulting theories should be of a general nature. When restricted to propositional languages, simple implicative formulas will do the job in both cases, provided there is enough room for three kinds of truth: propositions may be actually true, 'physically true' and logically true. One plausible way to do this is using a modal propositional language (Zwart, 1998), but there may well be other ways.

From the above sketch of possible connections, concluding with one example in some detail, it will be clear that there is not much competition between computational and classical philosophy of science, as soon as one is willing to reject the dogma of the impossibility of fruitful philosophical-methodological analysis of the Context of Discovery. But there is interesting possible competition between the computational philosophy of science and modern sociology of science, witness for example the title "Scientific discovery by computer as empirical refutation of the strong program" (Slezak, 1989). For, where is the dominant role of socio-cultural factors in computational discovery, evaluation and revision that many sociologists of science find so important? Besides such possibilities for competition, conciliatory positions are conceivable on both sides so that fruitful co-operation becomes possible. In particular, as mentioned in Subsection 11.1.2.4., Thagard proposes a few computer simulations of co-operation and competition between scientists that are intended to achieve sociological adequacy.

11.2. COMPUTATIONAL THEORY SELECTION AND THE EVALUATION MATRIX

Introduction

Four years after his pioneering book on the computer program PI (Processes of Induction) Computational Philosophy of Science (Thagard 1988), Paul Thagard published a new book entitled Conceptual Revolutions (Thagard 1992), on the ECHO program. In this last book he not only studies conceptual change in much more detail than before, he also replaces his 1988-approach to theory selection with a refined version.

Our claim in this section is that the architecture of theory selection in both cases seems to be on the wrong track. PI and ECHO impose a mixture of distinct considerations, such as explanatory value, simplicity, analogy, competition and the like. In our descriptive-cum-normative perspective, theory comparison is and should be essentially stratified, giving priority to explanatory evaluation. Hence, first the explanatory merits and problems should be assessed and compared. Only if that does not lead to the conclusion that one theory is explanatorily superior to the other, should additional considerations be brought into play in order to turn the scale, if one aims at selection at all in this case.

Thagard's primary aim is to design a set-up that is historically adequate, in the sense that simulation should lead to the same theory selection as scientists made in the past on the basis of the same elementary judgements. In his last book, Thagard presents in full detail five historically adequate selections made by the quasi-connectionist ECHO program: they concern the theories of Lavoisier, Darwin, along with three selections from geology. In this section we will show, by inserting the relevant input from Thagard in the simple Evaluation Matrix (EM) presented in Chapter 8, that our approach, i.e., explanatory merits first, also leads to the historically adequate selections, but now in a straightforward and transparent way. Hence, by way of meta-application of the stratified approach, we claim that our approach, being explanatorily as equally successful as Thagard's approach (as far as Thagard's explicit evidence is concerned), is to be preferred because of its simplicity.

In the last section we will discuss Thagard's reaction to the EM-approach, starting with Thagard's second aim of not only simulating the results of historical theory selection, but also the main processes, or at least the role of the main considerations involved, hence involving questions of psychological adequacy.

To be sure, this section is not a review of both of Thagard's books. For the 1988 book we refer to Subsection 11.1.2. Like that book, the 1992 book also presents an enormous number of fresh ideas and criticisms by looking at the philosophy of science from a computational point of view. However, our single aim in this section is to argue for a different approach to computational theory selection. In this respect, we want to illustrate the possibility of fruitful interaction between neo-classical and computational philosophy of science.

11.2.1. Unstratified approaches

In this subsection we will show in some detail that both PI and ECHO deal with evaluation in an unstratified way, such that explanatory superiority may be sacrificed to simplicity.

11.2.1.1. The unstratified architecture of PI

As we have seen, Thagard's 1988-approach to theory selection is essentially a kind of inference to the best explanation (IBE), where the best explanation is determined by three criteria: consilience, simplicity and analogy. The implementation in PI is restricted to consilience and simplicity, and the relevant fragment of PI is straightforwardly of a classically computational nature. With respect to PI we will now, in addition to Subsection 11.1.2.3., (only) argue that it is essentially unstratified, despite some formulations made by Thagard suggesting the contrary. In particular, PI may well sacrifice explanatory superiority to simplicity. Our descriptive stratification claim, however, is that scientists have almost never made such a sacrifice. To be sure, Thagard does not explicitly claim to have historical cases of this kind. Moreover, the actual historical cases discussed by Thagard (1988) are not fully worked out, a fact that makes them difficult to evaluate. Hence, our claim with respect to PI is only that as far as cases of explanatory superiority are concerned, PI is not only very complicated, but, assuming that our descriptive stratification claim is correct, it is also likely to lead in many cases to the historically wrong verdict. Note that we do not claim something for the case of (seriously) divided success. As far as divided explanatory success is concerned, PI may be on the right track.

Let us consider PI in some detail. The consilience of a hypothesis is defined as the sum of the relative importance, or weight, of its explanatory successes. The simplicity of a hypothesis is defined in such a way that it decreases when the number of auxiliary hypotheses required for the explanatory successes increases. PI selects the hypothesis with the highest value, defined as the product of simplicity and consilience. Of course, these arrangements are such that if both consilience and simplicity are favorable for the same hypothesis then that hypothesis is selected. However, consider the case that the consilience of one hypothesis is straightforwardly superior to that of another, in the sense that all explanatory successes of the latter are also explanatory successes of the former. Hence, in this case the sum of the weights of the successes of the latter does not exceed the sum of the weights of the successes of the former, whatever (positive) weights are assigned to specific successes. In other words, in this case the weights of the successes, of course a cause for debate, do not play a substantial role in the comparison of the consilience of the hypotheses. But the PI-arrangement is such that, if the hypothesis with higher consilience happens to be less simple than the other hypothesis, both the degree of simplicity and the total importance of the explananda come to play a substantial role. Hence, in that case it may occur that explanatory superiority is sacrificed to simplicity.

11.2.1.2. The theory of explanatory coherence (TEC) and the architecture of ECHO

In his 1992-approach to theory selection Thagard first develops a theory of explanatory coherence (TEC), governed by seven principles (pp. 65/66), "that establish relations of explanatory coherence and make possible an assessment of the acceptability of propositions in explanatory systems". The seven principles include: symmetry of coherence relations (P1), coherence obtained from explanatory relations, differentiated by simplicity considerations (P2), coherence obtained from analogical considerations (P3), 'incoherence' obtained from contradiction and competition (P5/6), acceptability of propositions due to data priority (P4) and due to coherence and explanatory impact (P7). According to TEC, a new theory will replace an old one if its hypotheses possess greater explanatory coherence. Though TEC, the underlying selection theory of ECHO, may be seen as a refinement of IBE, the underlying selection theory of PI, ECHO is essentially different from the relevant, classically computational, fragment of PI. Like many other fragments of PI and ECHO, ECHO implements TEC in a quasi-connectionist way by using the techniques of activation spreading and parallel constraint satisfaction. It is not straightforwardly connectionist, for each piece of data and knowledge is represented by a single node rather than distributed throughout the network. However this may be, as one may expect, 'ECHO-selection' is a non-transparent updating process, which may nevertheless lead, as a rule, to an unambiguous conclusion.

The input propositions are of the following kinds:
evidence-propositions, including 'negative' evidence-propositions, and hypotheses of a theory, including auxiliary hypotheses and additional facts. Input claims are made with input propositions, of which there are four types: explanation claims, analogy claims, contradiction claims, and data claims. Data claims only concern evidence-propositions leading to evidence statements, as we will call them, and additional facts. The other claims may concern all kinds of combinations of propositions. For example, not only the explanation of an evidence statement by certain hypotheses is taken into account, but also the explanation of a hypothesis by some other hypotheses, or by some background additional facts.

Starting with activation of the evidence-propositions and zero activation of all other propositions, after a number of updating runs all propositions approach their asymptotic activation value. Those getting a positive value, including the propositions with a data claim, constitute a coherent set, and they are accepted. Those getting a negative value are rejected.

Our main point of criticism of ECHO, to be elaborated in what follows, again concerns the unstratified interplay of selection considerations.

11.2.2. Priority of explanatory superiority (PES)

In this subsection we will first restate the basic features of the evaluation matrix, then show, using five examples, that it leads to the same results as ECHO, and hence that it deserves further consideration to be used in evaluation programs.

11.2.2.1 The evaluation matrix (EM)

In our approach explanation is also crucial, even though we will restrict it to the explanation of evidence statements by a theory. A theory is supposed to consist of all relevant hypotheses, including auxiliary hypotheses and additional facts. Except when otherwise stated, we will follow Thagard in using a liberal, not strictly deductive, notion of explanation, leaving room for all kinds of special types.

Recall that we introduced in Section 8.1. the (comparative) evaluation matrix (EM) in which two theories were compared in terms of their ability to account for the (individual or general) facts. Although the definitions of positive, negative and neutral (evidential) facts for a theory were essentially just (conditionally) deductive, nothing prevents us from the suggested more liberal interpretation of the logical relation on the one hand, combined with an explanatory interpretation on the other. Only when our further conclusions depend on the deductive relation do we have to make a proviso.

The matrix for theories of *X* and *Y* and a certain fact was as follows:

			X	
		negative	neutral	positive
	negative	B4:0	B2:-	B1:-
Y	neutral	B8:+	B5:0	B3:-
	positive	B9:+	B7:+	B6:0

The (comparative) evaluation matrix (EM)

From the perspective of *Y* the boxes B1/B2/B3 represent unfavorable facts, B4/B5/B6 indifferent facts, and B7/B8/B9 favorable facts. The numbering of the boxes was determined by three considerations: increasing number for increasingly favorable results for *Y*, symmetry with respect to the \-diagonal, and increasing number for increasingly positive indifferent facts.

The claim that *Y* is more successful than *X* in the light of the available facts was defined by: B1/2/3 empty (no unfavorable facts for *Y*) and at least one of B7/8/9 non-empty (some favorable facts for *Y*), with corresponding versions of the comparative success hypothesis and the rule of success, prescribing, for the time being, the choice of the theory that has so far proven to be more successful than others.

We defined the claim that *Y* is almost more successful than *X* in such a way that there is room for some 'Kuhn-loss': the new theory may no longer retain (or at least may no longer seem to retain) a success of the old one, without contradicting it (i.e., a B3-fact). The condition is that some (unfavorable) B3-facts are allowed, provided there are (favorable) B8- or B9-facts or the number of B3-facts is (much) smaller than that of their antipodes, that is, B7-facts. For the detailed motivation, including calling B7-facts weakly favorable for *Y* and B9-facts strongly favorable, see Section 8.1.

From now on we will interpret EM and the two comparative definitions as means of dealing with explanatory successes in a liberal sense. Since EM only takes explanatory considerations into account, and not simplicity, it may be seen as the elaboration of the idea of Priority of Explanatory Superiority (PES), in the same way as ECHO was the implementation of TEC.

In the next subsection we will apply EM to the five cases extensively treated in (Thagard 1992), where all formally indicated statements can easily be found.

11.2.2.2. Five applications of EM

Lavoisier

We will start with the relevant input used by Thagard for the Lavoisier example. The phlogiston theory (PT) as well as the oxygen theory (OT) is summarized by Thagard in six hypotheses. Eight evidence (E-)statements are considered. Seven of them can immediately be classified on the basis of the input claims as follows: 2xB6 (neutral) and 5xB7 (weakly OT-favorable). According to Thagard, these seven E-statements were the major E-statements taken into consideration by Lavoisier. Hence, from Lavoisier's point of view OT is explanatorily superior to PT.

For some reason or other, Lavoisier did not investigate OT's explanatory relation to the eighth E-statement (named E2 by Thagard): "Inflammability is transmittable from one body to another". Following Thagard's input that E2 is explained by PT, it belongs to one of the boxes B1/3/6, but he makes no mention of his opinion about OT in this respect. However, it seems plausible to assume that it would have been easy for Lavoisier to argue that OT is at least compatible with E2, hence that E2 should be placed in B3 or B6. In other words, for Lavoisier E2 would have been at most a case of Kuhn-loss, to be compared with 5 opposite B7-cases. Moreover, it even seems that it would have not been very difficult to argue that E2 can be explained by OT, and hence that E2 belongs to B6. In sum, according to Lavoisier's account, OT is explanatorily superior to FT, and at least almost, if we also take E2 into consideration.

Darwin

Next we will consider the Darwin example. Thagard discerns 14 evidence statements (E11 is missing for some reason or other). Darwin's theory (DT) is summarized in 3 main hypotheses, 3 auxiliary hypotheses, and 7 additional facts. These additional facts will not play any role in our evaluation. The creationist theory (CT) is summarized in the statement: "Species were separately created by God". Of the 14 evidence statements, 12 can be immediately classified on the basis of Thagard's input claims as follows: 3xB6 (neutral) and 9xB7 (weakly DT-favorable).

Although the presentation suggests that E6, "Once extinct, species do not reappear", should also be classified in the B7-box, one might claim that this statement is not only explained by Darwin's theory, but also by the creationist theory, at least by a deistic version. Hence it should be placed in box B6. But neither of the two alternatives changes the judgment that DT is explanatorily superior to CT, as long as we do not take E4 into consideration: "Species when crossed become sterile".

Thagard claims of course that E4 is explained by the creationist theory. However, Darwin is not explicitly reported to deal with E4 in the same way as with three other evidential statements explained by CT (E1,E2,E3) and also considered by Darwin as potential counter-evidence to DT. Darwin is able to explain these cases with his theory (using the auxiliary hypotheses). Hence, E4 seems to be of a similar nature as the exceptional case in Lavoisier's example. E4 is not really in conflict with DT, hence it should be placed in box B3 or B6. Further research might have led to the conclusion that it can be explained by DT (possibly with the aid of some extra auxiliary hypothesis), hence that it may be placed in B6, otherwise it is to be registered as a case of Kuhn-loss (B3), against 9 opposite B7-cases. In sum, the overall conclusion is essentially the same as for the Lavoisier example. According to Darwin's account, DT is explanatorily superior to CT, and at least almost, if we also take E4 into consideration.

Three geological cases

The third explicitly presented example by Thagard in fact concerns three different geological evaluations. Wegener's own evaluation of his theory of continental drift (WT) against the contractionist theory (CT), the same evaluation by his opponents, and finally Hess's evaluation of his later 'sea floor spreading' version of continental drift (ST) against CT. The evaluations produce both strongly favorable results (B9) and some negative evidence (NE-)statements as well.

Let us begin with Wegener. Thagard summarizes CT in 10 hypotheses, and WT in 12. It is important to note that Wegener borrows the so-called isostasy hypothesis (C8) from the contractionist, stating that "Large sections of the crust are in gravitational equilibrium". There are 20 evidence statements, plus two negative ones, NE6 and NE11, contradicting E6 and E11, respectively. All simply positive evidence statements, without corresponding negative ones, hence E1-E20, except E6 and E11, are easily classified: 9xB6 (neutral) and 9xB7 (weakly WT-favorable). Given that WT not only explains E6, but also that CT explains NE6, their combination E6/NE6 is a proper case of a strongly favorable result for WT (B9).

The situation is more complicated for E11/NE11. WT is simply reported to explain E11, but CT is not only said to explain NE11, but also to explain E11 (by some other hypotheses of CT, including the overlapping C8). Hence, the pair E11/NE11 may be placed in B6, B7 or even B9. However, none of these options influence the overall judgment that WT is explanatorily superior to CT.

Of the three options where to place E11/NE11, B7 seems the most reasonable for explaining E and NE, hence not-E, is comparable to explaining neither E nor not-E. In this case the final result is 9xB6, 10xB7 and 1xB9. Another possibility, would be to split the placement of E11 and NE11, in which case E11 would come in B6 and NE11 in B9. But this procedure seems unjustified for unambiguous cases. e.g., when CT would just explain NE11, for then it essentially leads to a doubling of the results. Hence, in the ambiguous case of opposite explanatory claims of one theory with respect to E and NE we prefer to say that the theory is indifferent to that couple.

In the case of the opponents, the two opposing theories are reformulated in the light of points made by Wegener's critics and, as one may expect, of course some other evidence statements are taken into consideration, 5 simple positive ones and 13 couples of positive and negative versions. Placing pairs in the way described above, we obtain the following overall result from the perspective of the contractionist theory: 1xB3 (weakly CT-unfavorable), 3xB6 (neutral) and 1xB7, 2xB8, and 11xB9 (weakly, moderately and strongly CT-favorable). Hence, apart from one case of Kuhn-loss (B3) to the effect that the contractionists do not claim to have an explanation for Wegener's primary observation: "The outlines of continents on the opposite side of the Atlantic are similar" (E3), their theory certainly is explanatorily superior to Wegener's from their evidential point of view.

Note that there occur two cases of B8, which had up to now not been instantiated: negative evidence explained by one theory (WT) and indifferent for the other (CT). Just to give an idea, one of these two statement pairs says "There are no mountains on the ocean floor" (E10), indifferent for CT, versus "Rigid ocean floors would also be compressed and form mountains" (NE10), predicted by WT.

Let us finally consider Hess's evaluation of his seafloor spreading version of continental drift (ST) as opposed to the contractionist theory (CT), summarized in 17 and 19 hypotheses, respectively. There are 24 evidence statements, 5 coupled with a negative evidence statement. There is one ambiguous couple, which is, for similar reasons as in earlier cases, placed in B7. The overall outcome, showing the explanatorily superiority of Hess's theory, is: 12xB6 (neutral), 8xB7 (weakly ST-favorable), 4xB9 (strongly ST-favorable).

11.2.2.3. Further elaboration and motivation

The general conclusion of the foregoing subsection is that the EM-simulation (by hand, hence by computer) of the five examples on the basis of PES is historically equally adequate as Thagard's ECHO-simulation based on TEC. Given that PES/EM is much simpler than TEC/ECHO, meta-application of PES/EM seems to prefer PES/EM, at least on the object-level. The point is not only that TEC/ECHO is much more fine-grained than necessary, its unstratified set-up is at least as problematic. Hence, there is sufficient reason to evaluate and elaborate (and program) the stratified PES/EM-approach further. To begin with the further evaluation, all four revolutions in physics informally dealt with by Thagard (Copernicus, Newton, relativity theory and quantum theory) seem cases of straightforward explanatory superiority. In the last section we will argue that this is actually the case for the examples of Copernicus and Newton.

With respect to further elaboration the natural question is, what to do in the case of equal or seriously divided explanatory success? One possibility, closest to TEC/ECHO, is to bring in other considerations, such as simplicity and analogy. Whereas we have doubts about any evaluative role of analogical considerations (of course, we are not against a heuristic role), we find it very plausible to bring in simplicity considerations, at least in the case of equal explanatory success. Suppose that on closer inspection in the explanatory evaluation of a theory one sub-hypothesis does not play any role at all. Then the theory obtained by omitting that sub-hypothesis, being equally successful as the original one, is to be preferred to the original one. This simplicity approach to equal success, may be extrapolated for divided success in one way or another, or it may be combined with the quantitative approach to divided success, as suggested below. (Of course, simplicity and analogy considerations may also be applied in cases of selection within a set of equivalent theories.)

Cases of divided success may be approached by some quantitative weighing of neutral and (un-)favorable results. For this purpose the reader is referred to Section 8.1., where a quantitative version of the evaluation matrix (QEM) was presented. QEM suggests, of course, that simplicity only comes into play in the case of quantitatively equal success, where the latter possibility is just a matter of sheer accident.

As far as programming is concerned, the implementation of PES/EM in the rest of ECHO (i.e., ECHO minus its specific TEC features) will be a relatively easy task. But PES/EM is not only a proposal for a local adaptation of ECHO, as already suggested in Subsection 11.1.4., we also conceive it as an explication of most of the explanatory and/or success evaluations built into other programs of theory selection and revision, in particular, most of those occurring in the collection edited by Shrager and Langley (1990) and dealt with in Subsection 11.1.3.

One reason to prefer the PES/EM-approach to the TEC/ECHO-approach was the way it deals with simplicity, the other reason is that PES/EM is directly suggested by our theory of truth approximation (ICR, Chapter 7). When restricting the attention to deductive explanatory relations, the core of PES/EM essentially corresponds to, a symmetric version of, the Rule of Success (RS), prescribing to choose (for the time being) the most successful theory. It has been demonstrated in ICR (Chapter 7) that RS, though not guaranteeing truth approximation, is nevertheless functional for truth approximation in a very transparent way, due to the non-trivial success theorem that being closer to the truth guarantees being at least as successful. To be sure, the theorem is almost trivial when no (relative) distinction between theoretical and observation terms is made. In that case it is a direct consequence of the definition of 'being closer (more similar) to the truth', of 'being more successful than' and the assumption that the data are correct. However, if there is a distinction between observational and theoretical terms the theorem is not at all trivial, and even needs some extra conditions. Recall that 'the truth' in the context of truth approximation is conceived as the most informative and, hence, strongest true theory that can be formulated with the available conceptual means. Here we confine ourselves to some specific remarks about PES/EM from this perspective.

First, it is important to note that PES/EM, like RS, can unproblematically deal with falsified theories, due to the room it allows for counter-evidence, i.e., boxes B1/2/4 for *Y*. To be precise, *Y* may well be preferable to *X*, even if *Y* has been falsified, provided these falsifications also apply to *X*, i.e., B4 is non-empty, but B1 and B2 are empty.

Second, this means that RS, as discussed in ICR (Section 7.5.), and more generally PES/EM, conceived as a rule of inference, is a severely corrected version of Inference to the best explanation (IBE). IBE should not be interpreted as the assertion that the best explanation is true, but it is the closest to the truth. In this way it cannot only deal with falsified hypotheses, but the conclusion is equally relative to the available explanations as the premises are. The standard version of IBE is in both respects rather problematic.

11.2.3. Process- or only product-simulation?

Paul Thagard, in correspondence on a previous version of this section, has noted that the EM-proposal is attractive, among other things, because it ties in with truthlikeness, but it "strikes as only an approximation to what is involved in theory selection. It ignores factors that scientific practice convinces me are important to theory choice: simplicity, analogy, and explanation of hypotheses by higher order hypotheses. The merit of ECHO is that it takes these into account AND explanatory breadth..." Moreover, Thagard adds the interesting remark: "The Newton and Copernicus simulations [by ECHO (Nowak and Thagard 1992 a/b)] are far larger than the ones in Conceptual Revolutions."

The following reply will illuminate the points of discussion. We do not want to dispute at all that factors like simplicity, analogy and higher order explanation prima facie play an important role in historical theory selection. These factors are frequently explicitly discussed. Hence, historical adequacy may be restricted to product-simulation, but one may also aim at a process-&-product-simulation of historical theory choices. In the latter case, these cognitive factors should be included. In that respect, TEC/ECHO may very well be on the right track. Although Thagard primarily stresses in Conceptual Revolutions that TEC/ECHO can reproduce the resulting choices, there is from time to time also the suggestion that he wants to be realistic with respect to the choice processes as well, and both the above quoted reaction and (Nowak and Thagard 1992 a/b) make this very clear.

Our claims are restricted to product-simulation, hence without the aim of psychological adequacy. As far as the five examples dealt with in Subsection 11.2.2.2. are concerned, it is possible to reproduce the resulting choices only on the basis of the reported explanatory successes and problems. Hence, the other reported factors are redundant insofar as product-simulation is concerned. A minor point of criticism of Thagard's 1992 work becomes relevant here. He calls theory selection an essentially holistic affair, as opposed to piecemeal theory formation. This view is misleading, at least as far as ECHO-selection is concerned, for it is evident that ECHO can make distinctions between the hypotheses of a theory. But Thagard does not report historical or artificial cases of such differentiated evaluation of the hypotheses of one theory. However this may be, it is clear that (implementation of) EM can be applied to subsets of the hypotheses of a theory, and hence also be challenged for reproduction of differentiated evaluation. As suggested before, in that case, simplicity will come into consideration: if an extra hypothesis does not change the success, it will and should be left out.

The Newton and Copernicus cases present interesting challenges to the main product-simulation claim. In the first place, as Thagard stresses in his quoted reaction, their ECHO-simulations are much more complicated than the five cases mentioned in the book. On the other hand, determination of the EM-choice, although laborious, can still be done by hand. The second, and most important challenge is of course whether the resulting choice is, like the ECHO-choice, in agreement with the historical choice.

Only 45 minutes of calculation by hand showed the following, using the inputs of Nowak and Thagard. Copernicus against Ptolemy leads to the results: 51xB6 (neutral), 8xB7 (weakly favorable) and 2xB9 (strongly favorable). Newton against Descartes to: 2xB3 (weakly unfavorable), 17xB6 (neutral), 2xB7 (weakly favorable), 9xB8 (moderately favorable) and 8xB9 (strongly favorable). Hence, the choices of Copernicus and Newton are also easily reproduced by the EM-approach.

In general, the challenge of new cases is that they may lead to strong counter-examples of the claim that the EM-method reproduces the historical choices: the EM-method might prescribe the opposite choice. If there are such cases our stratified model is descriptively inadequate, i.e., even with respect to the simulation of products. For Thagard (1992) such a counter-example to a general claim about historical choices would seem to be a too high price. However, following from Thagard's Chapter 7 of (1988), entitled "From the Descriptive to the Normative", we would seriously consider the possibility of occasional deviations from the norm. Moreover, we would also like to accept the extra challenge of trying to show for the particular counter-example that the historically made choice was less adequate from the truth approximation perspective. Of course, given that we never know what the relevant strongest true theory is, the latter evaluation will only be possible in terms of our favorite theory at the moment. However this may be, the extra challenge is still waiting for a good counter-example, i.e., the challenge of new cases.

The latest potential counter-example involves the wave versus particle theory of light. In response to Achinstein's opposite claim (Achinstein, 1990), Eliasmith and Thagard (1997) convincingly argue that in the nineteenth-century debate about the nature of light coherentist, and not probabilistic, arguments played a central role. They claim that their coherence account has two major advantages over Achinstein's probabilistic account: computational tractability and psychological plausibility. Now ECHO and EM both provide specifications of a coherence account. Assuming that the EM-version also leads to the choice of the wave theory, the relative merits of the three accounts can be measured in terms of these two criteria. Eliasmith and Thagard make it very clear that the ECHO-account is computationally much more tractable than a probabilistic account. However, it is also evident that, whatever the outcome, the EM-account is computationally still much more tractable than the ECHO-account. Similarly, Eliasmith and Thagard make it also clear that the ECHO-account is psychologically more plausible, referring to psychological evidence in favor of connectionist models of mind on the one hand and low probabilistic skills on the other. We would, however, like to claim that the EM-account may well be still much more plausible, provided the evidence points more or less unambiguously in one direction. The latter is indeed the case, that is, the wave theory is indeed explanatorily superior to the particle theory on the basis of the presented reconstruction.

According to the Appendix to Eliasmith and Thagard's paper, there are 23 evidence statements. Apart from the main statements, the two theories (WT and PT) require one or more additional hypotheses for the explanation of one or more evidence statements. They both need one extra hypothesis, not the same one, for one and the same evidence statement. Moreover, PT needs three more extra hypotheses, each for a different evidence statement. Hence, WT is evidently simpler than PT, and if WT and PT would explain the same facts, this would play an important role. However, the supposition is quite counterfactual for the score is as follows: 1 E-statement apparently belongs to B5, 11 to B6, 10 to B7 and 1 to B8. Hence, WT is explanatorily superior to PT, without any reservation.

In the present case, the authors do not argue extensively for the claim that ECHO provides a process-reconstruction; there is just one quote from Young in which he stresses the importance of explanatory coherence much more convincingly than that of simplicity. However, even if they would argue this convincingly, there remains the possibility that simplicity arguments in case of explanatory superiority just serve as additional considerations that might convince people in doubt to agree with the choice. Moreover, if scientists, like ECHO, really intermingle explanatory superiority with simplicity considerations, we may raise the normative question: is the actual process to be criticized for being not as simple as possible? If so, computer assisted research should of course ultimately be guided by that criticism and not by actual processes.

Concluding remarks

In the introduction to this chapter we mentioned four claims, viz., historical, psychological and philosophical adequacy, and practical relevance. In the last section it was shown that there may well be a tension between them. Even if ECHO is historically and psychologically adequate, it may well be that it is philosophically and practically questionable. On the other hand, it may well be that there are practically more useful alternatives that are philosophically adequate, historically adequate as far as product-simulation is concerned and that may even also be a first approximation of psychological processes.

To be sure, the possibility for the computational philosophy of science to be of considerable practical relevance is still far away. However, in principle the perspective of more or less standard computer assisted discovery, evaluation and revision need not remain science fiction. To begin with, Langley (1998) presents a systematic survey of seven specific examples of novel, computer-aided discoveries that have appeared in the scientific literature. They include a taxonomic discovery in astrophysics, a qualitative law in biochemical cancer research, a qualitative law in metallurgy, a quantitative relation in graph theory, a temporal law in ecology, a structural law in biochemistry, and, finally, a process law in catalytic chemistry. [5] Moreover, coupling uniformly represented knowledge bases from various fields would create an enormous general source for analogical reasoning and for producing surprising data on which we can proceed. In such specific and general ways the computational philosophy of science could join forces with the already flourishing practice of another type of computer assisted research, viz., computer simulation.

SUGGESTIONS FOR FURTHER READING

The suggestions for further reading mainly indicate the most important references in the text.

11 Computational philosophy of science

The main references in Section 11.1. were to Langley et al. (1987), Thagard (1988) and Shrager and Langley (1990). In Section 11.2. it was to Thagard (1992). For recent surveys, see Darden (1997), Langley (1998/2000) and Valdés-Pérez (1999).

EXERCISES

11 Computational philosophy of science

1) Look in your own field for a quantitative or qualitative law that might be (re-)discovered with one of the programs discussed in Subsection 11.1.1. Try to indicate the main heuristic operations that could be used.

2) Give examples in your own field of the four types of abduction indicated in Subsection 11.1.2.2. Give also an example of concept formation along the lines suggested there.

3) Discuss the pro's and con's of theory comparison according to PI (see Subsection 11.1.2.3. and 11.2.1.1.).

4) Look in your own field for an example of theory comparison with a generally accepted outcome in favor of one of them, and check whether PES/EM would lead to the same conclusion.

5) Paper-task: Check one of the mentioned examples of ECHO-comparisons 2.2.2 (Lavoisier, Darwin, Wegener), 2.3 (Copernicus, Newton, Huygens), that is, consult the reports of Thagard c.s. and check the reproducibility claim by PES/EM. Discuss the various possible aims of an evaluation computer program.

NOTES

CHAPTER 11

1. This chapter profited a lot from Alexander van den Bosch's detailed criticism. For a survey study of 'rationality in discovery', see Bosch (2001).
2. Davis' book is certainly not restricted to common sense knowledge, as the title of the book claims.
3. Chapter 3 of (Shrager and Langley, 1990) is, however, devoted to such problems.
4. As far as we know, there are only a few logical approaches to standard abduction, viz., the already mentioned belief revision developed by Peter Gärdenfors and others (Gärdenfors, 1988), non-monotonic logic approaches by Konolige (1990, 1996), logic programming (see Kakas et al., 1998), and the semantic tableau approach.
5. See Langley (2000) for an extended version. For some other recent surveys of approaches and results, see Darden (1997) and Valdés-Pérez (1999).

REFERENCES

Achinstein, P., (1990), "Hypotheses, probability, and waves", British Journal for the Philosophy of Science, **41**, 117-147.

Aliseda,A., (1997), Seeking Explanations: Abduction in Logic, Philosophy of Science and Artificial Intelligence, Dissertation Stanford, ILLC Dissertations Series (1997-04), Amsterdam.

Bechtel, W., (1988a), Philosophy of Science. An overview for cognitive science, Erlbaum, Hillsdale.

Bosch, A., van den, (2001), Rationality in Discovery. A study of logic, cognition, computation and neuropharmacology, PhD-dissertation, University of Groningen.

Darden, L., (1991), Theory change in science: strategies from Mendelian genetics, Oxford UP, Oxford.

Darden, L., (1997), "Recent work in computational scientific discovery", in: Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society, eds. M. Shafto and P. Langley, Lawrence Erlbaum, Hillsdale, pp. 161-166.

Davis, E., (1990), Representations of common sense knowledge, Kaufmann, San Mateo.

Eliasmith C., and Thagard P., (1997), "Waves, particles, and explanatory coherence", British Journal for the Philosophy of Science, **48**, 1-19.

Gärdenfors, P., (1988), Knowledge in flux: modeling the dynamics of epistemic states, MIT Press, Cambridge, Ma.

Holland, J., et al., (1986), Induction. Processes of inference, learning and discovery, MIT Press, Cambridge Ma.

Kakas, A., Kowalski, R., and Toni, F., "The role of abduction in logic programming", Handbook of Logic in Artificial Intelligence and Logic Programming, Vol.5, eds. D. Gabbay, C. Hogger, and J. Robinson, Oxford UP, Oxford, pp. 235-324.

Konolige, K., (1990), "A General Theory of Abduction", in: Automated Abduction, Working Notes. Spring Symposium Series of the AAA. Stanford University, Stanford, pp. 62-66.

Konolige, K., (1996), "Abductive Theories in Artificial Intelligence", in: Principles of Knowledge Representations, ed., G. Brewka, CSLI Publications. Stanford University, Stanford.

Kuipers, T., (1999b), "Abduction aiming at empirical progress or even at truth approximation", Scientific Discovery and Creativity, eds. J. Meheus and T. Nickless, special issue of Foundations of Science, Vol. **4**.3., 307-323.

Kuipers, T., (2000/ICR), From instrumentalism to constructive realism, Synthese Library 287, Kluwer AP, Dordrecht.

Lakatos, I., (1976), Proofs and refutations: the logic of mathematical discovery, Cambridge University Press, Cambridge.

Langley, P., (1998), "The computer-aided discovery of scientific knowledge", in: Discovery Science. First International Conference (1998, Fukuoka, Japan), eds. S. Arikawa and H. Motoda, Springer, Berlin, 1998, pp. 25-39.

Langley, P., (2000), "The computational support of scientific discovery", International Journal of Human-Computer Studies, **53**, 393-410.

Langley, P., Simon, H., Bradshaw, G., and Zytkow, J., (1987), Scientific discovery. Computational explorations of the creative mind, MIT Press, Cambridge.

Marek, W., and Truszczynski, M., (1993), Nonmonotonic logic: context-dependent reasoning, Springer, Berlin.

Mayer, M., and Pirri, F., (1993), "First Order Abduction via Tableau and Sequent Calculi", Bulletin of the IGPL 1: pp. 99-117.

Newell, A., and Simon, H.A., (1972), Human problem solving, Prentice-Hall, Englewood Cliffs, N.J.

Nowak, G., and Thagard, P., (1992a), "Newton, Descartes, and explanatory coherence", in: Philosophy of science, cognitive psychology and educational theory and practice, eds. R. Duschl and R. Hamilton, Albany: SUNY Press, pp. 99-115.
Nowak, G., and Thagard, P., (1992b), "Copernicus, Ptolemy, and explanatory coherence", in: Cognitive models of science, ed. R. Giere, Minneapolis: University of Minnesota Press, pp. 274-309.
Pirri, F., (1995), Abduction, PhD-dissertation, Université de Paris VI, Pierre et Marie Curie, Paris.
Selz, O., (1924), Die Gesetze der produktiven und reproduktiven Geistestätigkeit. Kurzfgefasste Darstellung, Cohen, Bonn.
Shrager, J., and Langley, P., eds., (1990), Computational models of scientific discovery and theory formation, Kaufmann, San Mateo.
Simon, H., (1969), "The architecture of complexity", in: The sciences of the artificial, MIT Press, Cambridge Ma., 84-118.
Simon, H., (1980), The sciences of the artificial, 2nd edition, MIT Press, Cambridge Ma.
Simon, H.A., and Groen, G.J., (1973), "Ramsey eliminability and the testability of scientific theories", The British Journal for the Philosophy of Science, **24**, 367-380.
Slezak, P., (1989), "Scientific discovery by computer as empirical refutation of the strong program", Social Studies of Science, **19.4**, 563-695, continued discussion in **21.1**, 1991, 143-156.
Smullyan, R., (1968), First order logic, Springer Verlag, Berlin.
Stillings, N., et al., (1987), Cognitive science. An introduction, MIT Press, Cambridge.
Tan, Y.-H., (1992), Non-monotonic reasoning: logical architecture and philosophical applications, Free University, Amsterdam.
Thagard, P., (1988), Computational philosophy of science, MIT Press, Cambridge.
Thagard, P., (1992), Conceptual revolutions, Princeton UP.
Valdés-Pérez, R., (1999), "Principles of human computer collaboration for knowledge discovery in science", Artificial Intelligence, **107** (2), 335-346.
Zwart, S.D., (1998), Approach to the truth. Verisimilitude and truthlikeness, dissertation Groningen, ILLC-Dissertation-Series-1998-02, 1998. Revised version to appear in the Synthese Library of Kluwer AP.