Rationality in Discovery

A Study of Logic, Cognition, Computation
and Neuropharmacology

Alexander van den Bosch

Rationality in Discovery

A Study of Logic, Cognition, Computation
and Neuropharmacology

ILLC Dissertation Series 2001-02

INSTITUTE FOR LOGIC, LANGUAGE AND COMPUTATION

For further information about ILLC-publications, please contact

Institute for Logic, Language and Computation
Universiteit van Amsterdam
Plantage Muidergracht 24
1018 TV Amsterdam
phone: +31-20-5256090
fax: +31-205255101
e-mail: illc@wins.uva.nl

RIJKSUNIVERSITEIT GRONINGEN

Rationality in Discovery

A Study of Logic, Cognition, Computation and Neuropharmacology

Proefschrift

ter verkrijging van het doctoraat in de
Wijsbegeerte
aan de Rijksuniversiteit Groningen
op het gezag van de
Rector Magnificus, dr. D.F.J. Bosscher,
in het openbaar te verdedigen op
donderdag 10 mei 2001
om 14.15 uur

door

Alexander Petrus Maria van den Bosch

geboren op 26 januari 1970
te Woudenberg

Promotores:	Prof. dr. Th.A.F. Kuipers Prof. dr. R. Vos
Beoordelingscommissie:	Prof. dr. E.C.W. Krabbe Prof. dr. P.G.M. Luiten Prof. dr. P. Thagard

opgedragen aan mijn ouders

School of Behavioral and
Cognitive Neurosciences

The investigations were supported by the Department of Philosophy of the Groningen University and the Groningen Graduate School for Behavioral and Cognitive Neurosciences (BCN). The printing of this thesis is further supported by the Institute for Language, Logic and Computation (ILLC).

ISBN 90-5776-062-2

Printed and bound by: PrintPartners Ipskamp B.V. Enschede, The Netherlands

This dissertation is also published electronically at: www.ub.rug.nl/eldoc/dis/fil/

Contents

Acknowledgements

Parts of the chapters of this thesis are based on papers that where published earlier:

Part I Introduction

Wetenschapsfilosofie. In Hendriks, P., Taatgen, N. & Andringa, T. Breinmakers en breinbrekers, Inleiding Cognitiewetenschap. Amsterdam: Addison Wesley Longman, Chapter 11, 1997.

Inference to the Best Manipulation - a case study of qualitative reasoning in neuropharmacy. In Foundations of Science 4 (4). Special issue on Scientific Discovery and Creativity: Case studies and computational approaches. Guest editors: J. Meheus & T. Nickles. p. 483-495, 1999.

Part II Discovery

Modeling Scientific Discovery in ACT-R. In J.A. Anderson (Ed.), Third annual ACT-R Workshop Proceedings, Pittsburgh: Department of Psychology, Carnegie Mellon University, 1996.

Discovering Patterns by Searching for Simplicity. In R. Valdez-Perez (Eds.), Systematic Methods of Scientific Discovery. Papers from the 1995 AAAI Spring Symposium (pp. 166-171). Menlo Park, California: The AAAI Press, 1995.

Part III Neuropharmacology

Rational Drug Design as Hypothesis Formation. In P. Weingartner, G. Schurz & G. Dorn (Ed.), 20th International Wittgenstein Symposium, I (pp. 102-108). Kirchberg am Wechsel (Aus): The Austrian Ludwig Wittgenstein Society, 1997.

Qualitative Drug Lead Discovery. In J. Meheus (Ed.). Working notes for the International Congress on Discovery and Creativity (pp. 163-165). Ghent: University of Ghent, 1998.

Met bijzondere dank aan:

Theo Kuipers voor zijn constructieve en kritische ondersteuning gedurende mijn tocht door de wetenschapsfilosofie; Rein Vos voor het mij op het spoor van de Parkinson casus brengen; Wia Timmerman voor haar tijd en bijdrage, en het tonen van haar werk in het lab; Ben Westerink voor het mij introduceren in de praktijk van de neurofarmacologie; Erik Krabbe voor zijn opmerkingen en discussie die ik altijd met plezier tegemoet zie; Niels Taatgen voor zijn reisleiderschap van mijn excursie door de ACT-R wereld; Hidde de Jong voor zijn nuttige commentaar op mijn werk; Jeroen van Maanen voor de geduldige uitleg en discussie over Kolmogorov complexiteit; I sincerly thank Paul Thagard for his comments, and I look forward to benefit from further discussion; Paul Luiten voor zijn verhelderende commentaar; Esther Stiekema voor de leuke gezamenlijke tijd bij het Filosofisch Instituut; Mark Kas voor de PowerMacintosh; Hauke de Vries voor het ondersteunen van deze en al mijn andere afwijkende hard- en software wensen in zijn netwerk; Knud Boucher voor het altijd paraat hebben van een origineel perspectief; Marco Hollenberg voor zijn morele steun en vriendschap; mijn ouders voor het hebben van meer geloof in mij dan ik zelf ooit zal hebben; Perrin en Myrthe voor hun onvoorwaardelijk dagelijks plezier in het leven; en tenslotte Petra voor haar ondersteuning op alle fronten en alle momenten.

Dank jullie wel.

Arnhem
March, 2001. Alexander van den Bosch

Part I Introduction

What is the rational use of theory and experiment in the process of scientific discovery, in theory and in practice? In this thesis I address this problem in three parts. I start with a general **introduction** (part I). Then I discuss three different theoretical models of the process of scientific **discovery** (part II). I finish this thesis with a discussion of a case study and model of discovery in the practice of **neuropharmacology** (part III).

In this first part I provide an introduction and overview of this thesis. I start with a specification of the **problem** (Chapter 1). Then as an introduction to part two I discuss some issues and views in the study of scientific **rationality** (Chapter 2). I finish this part with an introduction and overview of discovery in **neuropharmacology** (Chapter 3).

Chapter 1

Problem

1.1 Introduction

In this chapter I introduce the research problems about rationality in the process of scientific discovery that I faced in the years during my Ph.D. project and that are addressed in this thesis. To understand these subjects I studied different disciplines such as philosophy of science, cognitive psychology and artificial intelligence. An important problem for those disciplines is to understand what it means to be rational in the use and development of knowledge about the world.

It turns out that it is difficult to understand how we use common sense knowledge in everyday problems. I imagined that it would be less difficult to understand how scientific knowledge is used and developed. Common sense knowledge is almost by definition implicit, and therefore hard to understand. So, my idea was: why not concentrate on analyzing rationality in knowledge development that is supposed to be explicit, *i.e.* science?

So, I investigated different theories about rationality in discovery and the practice of discovery in neuropharmacology as a case study. This thesis presents the results of that investigation. I had to learn that while the product of scientific discovery is made explicit, the process of reasoning in the practice of discovery is often as implicit as common sense reasoning. So, I set myself the task to make it explicit, to understand rationality in discovery.

1.2 Goal

The specific goal of this thesis is to understand rationality in scientific discovery. Discovery is the act or process of making something known. Some scientific discoveries are made by accident, as a result of serendipity. But a goal of science is to make new discoveries by making use of theories and experiments to make things known. Theories are elaborate hypothetical assumptions, and experiments involve making specific observations of, and interventions in, natural phenomena. So, the method of scientific discovery is to make something known about the world with the use of theory and experiment.

To describe and understand the rationality in the process of scientific discovery I delve into the question of how acts in that process are suggested by reason. If an act is suggested by reason then there are arguments for doing something in that particular way, to achieve a particular goal. In sum, to understand rationality in scientific discovery I need to ask what it means to rationally use theory and experiment in the process of discovery in science.

1.3 Problem

To understand rationality in scientific discovery I analyzed different theories about the rational use of theory and experiment in science, and the practice of drug research for Parkinson's disease as a case study. An answer to the following specific problem is pursued:

Problem What is the rational use of theory and experiment in the process of scientific discovery, in theory and in the practice of drug research for Parkinson's disease?

An answer to this problem should provide an answer to the following specific questions about empirical science in general and neuropharmacology in particular:

Question 1 What is the structure of a scientific theory? Generally this question treats properties of scientific laws, theories and concepts. I will pursue this question in general and particularly for the dopamine theory of Parkinson's disease and related biological theories and concepts.

Question 2 What is the process of scientific reasoning? Traditionally, this question is about inference in the explanation and prediction of phenomena. I will also treat reasoning in the formation and revision of hypotheses.

Question 3 What is the route between theory and experiment? This question is relevant for understanding discovery in empirical science in general and drug research in particular. Not only do I investigate how the results of experiments influence theory, but also how theory and known (drug) interventions direct the suggestion for experiments.

1.4 Method

To pursue the problem of this thesis I undertook the following tasks:

- A survey of contributions to the study of scientific rationality (Chapter 2)
- A conceptual study of models of discovery in science as proposed in studies of logic, cognition and computation (Chapters 4, 5, 6)

- A review of the literature on the brain dopamine theory of Parkinson's disease (Chapter 7).

- A case study of the practice of experimental drug and brain research at the Pharmacy Department of the Groningen University (Chapter 8)

- Modeling the dopamine theory and the studied practice of discovery (Chapter 9)

- Summarizing the results of the case study of neuropharmacology (Chapter 3)

1.5 Background

The structure of theories and processes of scientific reasoning are investigated normatively in logic and artificial intelligence, and descriptively in cognitive psychology. In studies of logic scientific reasoning is mainly explicated as valid deduction of consequences. Studies of both artificial intelligence and cognitive psychology understand the process of scientific discovery as a kind of human problem solving. In that view it is held that human beings can solve scientific problems because they can (learn to) manipulate symbols.

The work of Newell and Simon (1972) sees the process of problem solving essentially as a search process, based on the manipulation of symbols. They defend the idea that for a problem it is possible to define a space of possible solutions that can be searched. This search is done by heuristic rules that, given the problem (the start condition), test whether the solution (goal condition) is being approached, and adjust the search accordingly.

In both artificial intelligence and cognitive psychology this process is investigated and modeled computationally. In order to do so it is necessary that the structure of the problem and the required knowledge is made explicit in a symbolic representation. Based on that representation, heuristic search rules must be able to effectively test if the goal state is being approached and if it is rational to pursue a particular direction of search.

John Anderson (1993) proposed a unified theory of learning and problem solving to explain rational behavior. This theory contains assumptions about the nature of explicit symbolic processes of reasoning, together with assumptions about implicit statistical processes of learning.

In another discipline, that of machine learning, one approach takes effective learning as searching the shortest computer program that can describe and predict patterns in observations (Li & Vitanyi, 1994).

In the third part of this thesis I shall investigate how the rational use of theory and experiment in drug research can be seen in terms of the role a theory plays in directing the search for an experimental drug intervention. A theory can be seen as a constraint in the search space of conceptually possible interventions. The main goal of drug research is to find an intervention that satisfies given conditions best. Testing a theoretically suggested intervention experimentally can either lead to a new drug or a new theory.

To compare theories about discovery, as set out in Part II, to scientific practice I will analyze the structure of problems in neuropharmacology in Part III, modeling the process of reasoning in rational search tasks with different kinds of goals.

The discussed models of problems in neuropharmacology will be based on the work of Benjamin Kuipers, Peter Karp, Theo Kuipers and Rein Vos. Benjamin Kuipers (1994) investigated how to represent qualitative knowledge about dynamical systems as qualitative differential equations, and how to reason with them correctly. Peter Karp (1992) made a computational analysis of the structure of molecular biological knowledge and the process of hypothesis formation in biological research. Theo Kuipers (2000) investigates logical structures and heuristic patterns in scientific research and Rein Vos (1991) investigates the logic of development in drug research. They explicated the development of drugs theoretically as a systematic attempt to bring together the properties of available materials and wishes for functional properties. In discussing neuropharmacology I will describe how biological theory can be used to infer those desired properties, to infer the best intervention.

1.6 Overview

This section gives a short overview of the subjects and problems that are discussed and the particular questions that are answered in the other chapters of this thesis.

Part I Introduction

The general problem of this thesis is: what is the rational use of theory and experiment in the process of scientific discovery, in theory and in practice? Part I discusses: issues in the study of rationality (Chapter 2), as an introduction to Part II; and my case study of neuropharmacology (Chapter 3), as an introduction to Part III.

Chapter 2 Rationality

This chapter provides an introduction to the discussion of discovery in Part II of this thesis. In that part I delve into ideas from cognitive psychology to look at issues about rationality in science that are traditionally part of the problems of the philosophy of science. The particular question that is answered in this chapter is: how can cognitive psychology contribute to the discussion about the rationality of science in the philosophy of science?

I argue that ideas from cognitive psychology in general can make a sensible contribution to debates about the rationality of science in philosophy. I make this point clear by explicating some relations between assumptions in cognitive psychology and issues in the philosophy of science.

Chapter 3 Neuropharmacology

This chapter is an introduction to my case study of neuropharmacology in Part III of this thesis. The particular question that is answered is: what is the rational use of theory and experiment in neuropharmacology?

This question is answered more extensively in Part III. I argue how the rational use of neurophysiological models can be modeled as goal directed reasoning about qualitative differential equations. To understand reasoning in neuropharmacology I distinguish inference to the best intervention from inference to the best explanation. I further briefly discuss how qualitative reasoning about neurophysiological models as part of a computer supported discovery system could aid in using, understanding, and testing models about large biological systems.

Part II Discovery

The specific problem of Part II (Discovery) is: what is the rational use of theory and experiment in the process of scientific discovery, in theory? This part discusses models of scientific discovery according to studies of: logic (Chapter 4); cognition (Chapter 5); and computation (Chapter 6).

Chapter 4 Logic

What is rationality in discovery, according to the study of logic? Traditional philosophers of science are usually interested in what scientific discovery ought to be, and how reasoning in that process can be valid or justified. I discuss how rationality in discovery is logically understood as valid reasoning, part of a circular process of observing, describing, explaining, predicting and intervening in natural phenomena. The particular questions that are answered in this chapter are: what is a scientific theory and what is scientific reasoning, according to the study of logic?

To address these questions I discuss an illustrative example of an explanation that contains a series of inferences that can be marked as fallacies from the viewpoint of logic. Yet, I argue that these inferences are common in science and part of abductive inference as defined by C.S. Peirce. I further make a categorical distinction between semantic abduction and material abduction. I argue how material abduction, together with other types of inductive inference, constitutes a part of semantic abduction. I conclude by answering the three specific questions (from section 1.3) of this thesis, from a logical point of view.

Chapter 5 Cognition

What is rationality in discovery, according to the psychological study of cognition? In cognitive psychology, rationality in scientific discovery is being studied as an interesting cognitive phenomenon, to be studied empirically. ACT-R is the name of a unified computational theory of cognition that aims to explain the data from studies of cognition. The particular question that is answered in this chapter is: how to understand and model scientific discovery with ACT-R?

I show and argue how the ACT-R model can learn by analogy the processes of two other cognitive models of discovery, called BACON and PI. I further discuss the nature of theory and method in the different cognitive models, and the difference between the logical and psychological views on explanation and prediction. I discuss how human performance on the Wason card selection task (an often performed psychological experiment where subjects test a hypothesis) seems irrational from a logical point of view. I propose a statistical model that can demonstrate the opposite. I conclude by answering the three specific questions of this thesis, according to the psychological study of cognition.

Chapter 6 Computation

Both the logical and the cognitive models of scientific discovery I discussed in the former chapters include a condition to prefer simple explanations. Yet these models do not show why it is rational to prefer simplicity. In Chapter 5 I discussed how the ACT-R model of cognition prefers simplicity as a consequence of a mechanism that prefers high probability. In this chapter I investigate the relation between probability and simplicity in the computational description, explanation and prediction of empirical data. The particular questions that are answered in this chapter are: how can simplicity most generally be defined and why should a simpler theory be prefered above a more complex one?

I discuss how the Minimum Description Length principle subsumes other definitions of simplicity and how the simplicity of a hypothesis can be related to the probability of its predictions. I conclude by answering the three specific questions of this thesis, according to the study of computation.

Part III Neuropharmacology

The specific problem of Part III is: what is the rational use of theory and experiment in the process of scientific discovery, in practice? This part discusses and models my case study of drug research for Parkinson's disease, *i.e.* I investigate: how Parkinson's disease and the effect of known drugs are explained by the dopamine theory (Chapter 7); the use of theory and experiments in a practice (Chapter 8); and a computational model of both the dopamine theory and the studied practice of discovery (Chapter 9)

Chapter 7 Theory

How are theory and experiments used in the practice of drug research for Parkinson's disease? To be able to address this problem I first survey the literature on the dopamine theory of Parkinson's disease. The particular question that is answered in this chapter is: how are Parkinson's disease and the effect of known drugs explained by theory?

I first provide a general introduction to Parkinson's disease. I then go into the basics of the dopaminergic nerve cell and the basal ganglia, which is the neural structure in the brain that partly controls voluntary movement, and how a defect in it causes Parkinsonian symptoms. I end this survey with a short overview and explanation of a selection of therapeutic drug interventions for Parkinson's disease.

Chapter 8 Practice

How are theory and experiment used in the practice of drug research for Parkinson's disease? In this chapter I report on my interviews with researchers at the Pharmacy Department of the Groningen University. These interviews where partly conducted while witnessing their work in the laboratory.

Several techniques are being used to search for new drugs and explore the activity of the basal ganglia. The particular questions that are answered in this chapter include: how are new drugs investigated and how are experiments being used to explore and test both new drugs and assumptions about the mechanisms of the brain?

Chapter 9 Discovery

In this final chapter I aim to explicate rationality in the process of discovery in neuropharmacology by describing both the theory and the studied practice, using the concepts from my theoretical discussion of discovery in Part II. The particular question that is answered in this chapter is: what is rationality in discovery in the case of neuropharmacology? First I discuss the use of models to describe theories about dynamical systems. Next I describe the structure of the dopamine theory of Parkinson's disease based on those models. By analyzing the reports about the practice of neuropharmacology I explicate a number of different routes between theory and experiment. I continue with a discussion of computational models of reasoning and discovery in biology. I conclude this chapter by summarizing my answers to the three specific questions of this thesis in the case of neuropharmacology. I then conclude this thesis by arguing that an answer to these questions can contribute to understanding rationality in discovery, as well as contribute to the process of scientific discovery itself.

*

Chapter 2

Rationality

2.1 Introduction

What is the rational use of theory and experiment in scientific discovery, in theory? In pursuing an answer to this problem in this thesis I use ideas from cognitive psychology to look at issues about the rationality of science that are traditionally part of the problems of the philosophy of science. As an introduction to that approach I will argue in this chapter how ideas from cognitive psychology can make a sensible contribution to debates about rationality in philosophy of science. I will make this point clear by explicating some relations between assumptions in cognitive psychology and issues in the philosophy of science. Hence, the particular question in this chapter is: how can cognitive psychology contribute to the discussion about the rationality of science in the philosophy of science? As an introduction to this chapter I first want to tell a parable about an intriguing family.

Prelude

This tale begins in the seventeenth century, in the days of the first great achievements of an ambitious young child of Mother Philosophy. In his time this aspiring infant developed a successful new style of understanding the world by tampering with it. He and his family became known as Experimental Philosophy.

Eventually he left the skirts of proud Mother Philosophy as an independent adolescent named Science. He set out to find answers to millennia old questions of the Philosophy family in a way that proved successful: with the method of empirical experiment and with the aid of the rigidity of his other parent, Father Mathematics. Science had an older brother who was fascinated by his younger brother's doings. When Science left Mother Philosophy his older brother stayed safely with his mother. Today the family of this son is called: Philosophy of Science.

At the end of the nineteenth century Mother Philosophy gave birth to a startling bright son from Father Mathematics. He was nurtured and raised by his brothers Philosophy of Mind and Philosophy of Language. With a few growing pains it reached maturity very quickly. The career of this son looked really promising. He begot the name Modern Logic. The Philosophy of Science family was very impressed by the capacity of this new sibling. Especially Positivism, a relatively young member of the Philosophy of Science family, was delighted.

At a certain moment in the beginning of the twentieth century, Positivism with the aid of Modern Logic declared it was time to completely break with one of Mother Philosophy's traditions. He thought that his admired brother Science had proved that the ideas of Mother Philosophy that where not empirical were irrelevant for answering Science's questions. All questions worth considering should be questions only Science could answer. And all theories of Science should be certain. Positivism would try to achieve this through the interpretation of the language of Science with the tools of his brilliant young brother Modern Logic. No wonder, he became known as Logical Positivism.

First this offspring of Philosophy of Science was warmly welcomed and embraced by old brother Science. It looked as if he opened up the possibility that the activities of Science would finally provide certainty, without being bothered by the problems and questions of Mother Philosophy that could not be solved empirically. Determined Logical Positivism would give Science a totally empirical and mathematical guideline and justification for his actions, in the true spirit of Science himself.

Yet the claims of Logical Positivism did not last very long without reaction of his brothers. Rationalism, another smart protégé of Philosophy of Science, showed, also with the help of his brother Modern Logic, that Logical Positivism's method of justification of the actions of Science was logically incorrect and he replaced it with an alternative. With this other method of Modern Logic as his standard he enforced the claims of the family of Realism, another son of Philosophy of Science. Logical Positivism felt defeated and eventually retreated from his too optimistic ideas.

Meanwhile, Science himself did not leave it with that. He took the discussion to his own domain to study the problems further. Science set out to empirically investigate his own activities and successes, those of that day and those of the glorious past. By empirically studying their own behavior, a member of the family of Science, Social Science, examined how the Sciences had actually behaved and tried to find out why they had done it in that way.

An astonishing conclusion seemed that a fully rational justification and explanation of Science's action and success seemed not possible. It appeared that beliefs of the members of the Science family were not rationally determined but socially. It seemed to have nothing to do with truth, the hallmark of Mother Philosophy as well as that of old Science. But, ironically enough, the truth of that conclusion implied that that conclusion could not be justified as true either. This was honey for the taste of a black sheep of the Philosophy of Science family: Relativism. According to him truth was impossible to achieve by Philosophy as well as by Science.

But the noble children of Philosophy of Science soon recovered from the apparent blow dealt to them by this relativistic conclusion. They put forward competing theories of rational justification of the activity of Science and tried to live up to Science's empirical standards.

Realism refined his ideas and, with the help of Modern Logic, delivered a rational justification of the data of Social Science, interpreted as steps toward the truth. Pragmatism, another member of the Philosophy of Science family, sought a way in the middle. He put forward a justification of the behavior of Science as rational, with not truth but just the solving of problems as his goal.

Looking in perspective to the quarrel among the Philosophy of Science family about the career of their successful brother Science we see so far the following: Logical Positivism tried to provide a justification of the results of Science's activities with the help of the formal language of Modern Logic, in which the Science family could represent their theories rigidly. Realism agitated against this by rejecting Logical Positivism's ideas about method and justification and put forward another method and way of justification, based also on Modern Logic, in its place. With the help of an empirical investigation of Science, Relativism reacted to Realism by showing that Science did not and never had acted according to Realism's rational method based on Modern Logic. From this it would follow that the doings of Science could not be explained rationally in this way. Realism refined his rational theory of justification showing that it could fit in with Science's empirical data. Next to that, Pragmatism developed another justification of Science's deeds by characterizing rationality as practical problem solving instead of looking for truth.

In the meantime, before Social Science studied the behavior of the members of his family as a group, other members of the Science family had not relinquished their attention to the subjects of behavior and rationality. Illuminated by the views of Logical Positivism, Behaviorism, a pretentious newborn of Science, studied behavior according to Positivism's methods and tried to keep rationality out of his theories. But Behaviorism disappeared to the back stage shortly after the rationalistic defeat of Logical Positivism. In his place came Cognitive Psychology, a new protégé of Science, who set out to explain behavior as a result of rationality. He did this with the help of new developments of Modern Logic and with the use of empirical methods. This ambitious child of Modern Logic and Science gave rise to a new style of thought for Philosophy of Mind. His family got the name Cognitive Science.

Until fairly recently the development of the family of Cognitive Science is looked at by Philosophy of Science as mainly another Science to be considered true, relative, practical or just nonsense according to some style of justification. Yet, Cognitive Science had the ambition of explaining the whole notion of belief and reasoning by his empirical examination of thought, rationality, the brain and behavior. Among other studies he did so by the exploration of Science's presumed mental processes during the process of discovery. He got precious help from Uncle Technology together with the ideas of the daughters and sons of noble Modern Logic and Science: Computer Science. Today Cognitive Science is supported with its work by a great part of the Science family.

I now come close to the moral of this introductory story. When we return to the entanglements of the Philosophy of Science family we noticed that at one point the relation between today's Science and Philosophy of Science turned around. Scientific results justified the claims of a Philosophy of Science instead of, in the traditional way, the other way around.

So, I ask: why shouldn't Cognitive Science contribute to Philosophy of Science's family discussions about rationality, reality and truth just as well as Social Science does? For the family of Science as well as for the family of Philosophy of Science it is probably insightful and productive if both directions are explored, possibly to result in a better relationship of understanding between the families…

Overview

In this chapter I discuss how ideas from cognitive psychology could be relevant in the domain of philosophy of science and where they would clash. After that I look the other way around, showing that cognitive psychology addresses topics relevant for philosophy of science that are usually not addressed in the mainstream.

In Section 2.2 I sketch some main issues in the philosophy of science, briefly rehearsing the ideas of several philosophers who made influential contributions to the field: Carnap, Popper, Kuhn, Lakatos, Laudan and Hacking, followed, in Section 2.3 by Jerry Fodor's thesis about the language of thought, a paradigmatic theory in cognitive science.

Then, in Section 2.4 I will consider the appropriateness and possibility of a contribution from the viewpoint of Fodor's cognitive psychology to the issues of rational justification, theory development, representation and the explanation of behavior. I will argue that if Fodor's thesis is accepted, no relevant form of relativism is tenable. His thesis provides a rich framework for considering theories about reasoning in a scientific context.

Following that, in Section 2.5, I shall argue that cognitive psychology can do more than that. It also makes possible a theory about discovery and it shows a relation with discovery as problem solving and the justification of theories. As a result of that, I will argue that the framework of cognitive psychology is rich enough to provide an adequate explanation of the development of scientific theories. In Section 2.6 I conclude this chapter with an evaluation of this chapter's claim, that accepting a theory about science is accepting a theory about the mind and vice versa.

2.2 Philosophy

There are many ways to look at science from a philosophical stance. There are probably just as many ways to describe existing philosophies of science. For that reason, as a guideline for this section I use the ideas of some of the important contributors to the field that is called philosophy of science. I follow four general issues in the work of these philosophers: theories about the justification of theories; the development of theories; what theories represent; and finally the actual practice of scientists. This line of description probably does not do complete justice to the initial intentions of these philosophers but is still useful for a short overview in the light of the goal of this chapter.

Carnap

After the sudden flourishing of theories in logic at the end of the last century, philosophy of science became focused on theories about language. Fast developments in logic provided formal languages with the aim to interpret propositions non-ambiguously. If it would be possible to interpret all theories of science in a formal language then the meaning of those theories would be reduced to the relation of the formal language and the world. This assumption gave rise logical positivism.

The main question became: what should or does language represent? Rudolf Carnap (Carnap, 1967) defended the following idea: all that a scientific theory should represent were terms and propositions whereof the truth could be confirmed in reality. The meaning of a proposition should be its way of verification.

Theory development should be a process of putting theories forward and confirming them. When universal statements were constantly confirmed they were justified by induction: justification of the general by the special. All scientific questions should be answered in this way to have any meaning at all. Questions which could only have answers that could not be confirmed should be dispelled from the domain of science. Theoretical terms of theories should be translatable into observational terms to be allowed in a scientific theory. Causes, unobservable entities, untestable hypotheses were all considered to belong to metaphysics and should have no place in a decent philosophy of science.

Popper

A philosopher who agitated most strongly against the logical positivism of Rudolf Carnap was Karl Popper. In his famous book 'The Logic of Scientific Discovery' (Popper, 1959, first published as 'Logic der Forschung' in 1934), he reacted with a logical critique against Carnap's ideas about representation, theory development and justification. He argued that confirmation as justification of universal statements is not tenable. A theory could be confirmed many times and still be false. For a theory to have any value it should be possible to refute it deductively.

Popper defended this notion as a way to demarcate true science from pseudo-science: a rational criterion of demarcation. The more sorts of experiment a theory allows to test it with, the more empirical content it has. The more falsifying tests it survives the more it is corroborated. A theory can never be accepted as true but only be found false as one accepts a refuting singular statement or otherwise be highly corroborated as one has come to accept many singular statements that support it.

The method of doing science should be to put theories forward and then to try to falsify them. When a theory is falsified it should not be repaired with ad hoc hypotheses. Only in this way, which became known as critical rationalism, could science produce justified knowledge about the world.

Just as the logical positivists, Popper thought that the discovery of theories was not a matter of logic. For understanding the logic of science it did not matter how a theory or law was discovered. That should be the concern of psychology. It did matter whether a theory could be tested and evaluated. That was purely a matter of logic, certainly not of psychology.

The only logic to discovery is that it can be validly discovered that a theory is false if a prediction of that theory is found to be false. So Popper's book could just as well been titled 'The Logic of Scientific Evaluation' because as far as Popper is concerned in his book, there is no logic of scientific discovery.

Kuhn

Thomas Kuhn looks at science from another perspective (Kuhn, 1970). Instead of thinking about what scientific theories should look like and how they should be developed, he empirically and historically investigated what the actual practice of Science did look like and had looked like.

What became clear to Kuhn was that scientists do not throw away their theories when they encounter a refuting counterinstance. They stick to them as long as possible. He noticed that science knows periods of normal science, in which puzzles are solved within the borders of a theoretical paradigm, next to periods of revolution. During a revolution the old theoretical paradigm is substituted by a new paradigm with different theoretical presumptions.

These historical and sociological facts looked nothing like Popper's story about the rational criticism of falsification. Most theories are born refuted and nevertheless function as the theoretical assumptions within paradigms. And when a new paradigm is accepted it looks as if there is no rational ground for it that has anything to do with the truth of the theories in the paradigms. So, real scientific practice seemed nothing like a rational affair in Popper's sense.

According to this fact it was argued that the meaning of the terms of a theory changed radically even if the same names were used in the new paradigm. That made 'truth' relative to a paradigm, what implied that there is no progress in science but merely succession. For, how can a paradigm say anything truthfully about the world when it is a matter of time for it to be rejected and succeeded by a radically new one?

So in sum, from empirical research in sociology and historical analysis of scientific development it follows that if 'truth' is regarded as a feature of a theory which is not falsified, and 'scientific rationality' amounts to rejecting a theory as soon as a falsification occurs, then science has nothing to do with truth and scientific practice has nothing to do with rationality. It was thought that beliefs are socially determined, not rationally, dependent on scientists' authority and social influences.

Lakatos

Imre Lakatos was strongly opposed to this irrational and relativistic picture of science (Lakatos, 1978). He built on Popper's ideas trying to show that they could be made consistent with the empirical data of the historical and social studies. He elucidated the activity of science not as the project of trying to refute one theory but as investigating empirical phenomena within the theoretical frame of a research program.

A research program consists of a theoretical core which is protected by a belt of auxiliary hypotheses. When a seemingly refuting instance is encountered an auxiliary hypothesis should be reconsidered, not the theoretical hard core of the research program. So in this way it can be explained why a theory does not get abandoned after falsification: one diverts the falsification to a protective auxiliary hypothesis. A revolution is explained as a change in the theoretical hard core.

What makes Lakatos' research programs really different from Kuhn's paradigms is that there is a rational way of determining when to give up on a particular theory. Lakatos evaluates progression in theory development on the basis of the increase of empirical content. The empirical content of a theory contains the possible models of a domain that are excluded by that theory: the higher the content, the more tests are possible to refute the theory. When a theory stops incorporating new facts, the program can be considered as degenerating. It then can be abandoned for any theory that does succeed in explaining those facts. So with this account it is possible to defend a notion of truth next to a rational justification of scientific theories and practice.

Laudan

In contrast to Lakatos' realism and research programs stands Larry Laudan's pragmatism and notion of research traditions (Laudan, 1978). A relevant difference with Lakatos' research programs is that, not the empirical content but a research tradition's ability to solve problems is central for the tradition's progress.

A theory is not considered as good at solving problems when it is progressive but the other way around: it is progressive when it is good at solving problems. Those problems range from logical inconsistencies and empirical problems to conceptual differences in the worldviews of scientists.

The rational choice between two theories in this way is for the theory which solves most problems. A notion of truth is not considered as relevant to judge scientific activity as rational. In this way scientific theories, and especially terms about non-observable entities within them, do not have to say anything about reality at all to be successful.

Hacking

Ian Hacking takes a different approach in the debate about truth (Hacking, 1983). He emphasizes the relevance of the practice of experiments in science. He argues that the philosophy of science is too much concerned with theory.

He accepts the reality of some theoretical entities but without accepting that a theory that explains a phenomenon must be true. You could establish the existence of, for example, electrons by doing intervening experiments, which result should be seen apart from the question whether the theory you test about electrons says anything true about reality. But from a rationalist viewpoint you can see the ability of intervention as just a disguised form of rational justification: accept a theoretical entity when it explains phenomena during intervening experiments.

Summary

You could frame the above theories about science as consisting of an opinion about: the behavior of scientists, their organization and their acts of adhering to or working on the basis of some scientific theory; what a theory represents, in other words its relation with reality; how a theory develops; and finally, how a theory is justified. So I can give the following summary.

In Carnap's view, a theory about science is first of all a theory about representation. All possible theories of science should exist of terms which refer to the observable world. Theories that are repeatedly confirmed are justified by induction. For Popper not all terms of a theory have to be observational. A theory that explains a phenomenon must be falsifiable through an experimental result that is implied by the theory. You can justify a theory rationally if it is not (yet) falsified. Kuhn stresses the presence of paradigms and revolutionary changes in science, implying that a theory never represents anything truthfully about the world that can be defended by Popper's critical rationalism. Lakatos tries to save truth by seeing theory development as research programs with an inviolable theoretical core that is protected by a belt of auxiliary hypotheses. A research program can rationally be abandoned when it stops explaining new facts while another research program can. Laudan's research traditions are considered progressive when they solve many and new problems, which is their goal and not the pursuit of truth. Finally, Hacking sees the truth of theories as a question other then that of the existence of theoretical entities, the latter can be established by explaining phenomena during intervening experiments by their most likely cause.

What all philosophers in this tradition have in common is that they do not attempt a further clarification of the role of processes of the mind of persons involved in science. Popper rejects psychologism, yet a form of psychologism that is based on psychological behaviorism. Kuhn embraces a psychology that implies multi-rationality but does not explain how it does so. Lakatos argues about why and when psychology could or should not interfere with the explanation of science, but he judges rationality irrelevant for it. For Laudan, science is problem solving. But he does not tell how that process comes about. Hacking also does not address the role of the mind in science (at least, not in the literature I reviewed).

In the following section I will explicate a general idea about cognitive psychology of Jerry Fodor's. It provides a framework for explaining and empirically investigating rationality in cognitive process. In Sections 2.4 and 2.5, the relevance of such a framework for the above ideas about science will be discussed.

2.3 Psychology

In this section I describe the general frame of assumptions Fodor's about cognitive psychology which is representative of the symbolic approaches in cognitive science. Fodor is a philosopher who contemplated that fundamental assumptions of cognitive psychology. In chapter 5 I will discuss work of the psychologist John Anderson, who aims to provide explanations for empirical data from psychological experiments.

One could characterize the program of cognitive psychology as looking for an explanation of intentional human behavior. The program grew out of the failure of behaviorism to explain the total scope of behavior, humans as well as animals, as a function of the environment. Cognitive psychology postulated again beliefs and desires in the organism in order to explain behavior that was judged intentional. The notions of logic and computation became recognized as a new way to accurately study language and thought. The mind of human beings was to be understood as a symbol manipulation system that governed all aspects that had made humans rational. Empirical data about complex behavior, thought and language could all be explained as the result of a process of symbolic computation that somehow should take place within the brain. .

What is now generally assumed in the program of cognitive science, is that cognitive processes of higher organisms should be seen as computational. Cognitive (or, as it is also called: computational) psychology made it possible to study language and the processes of thought with mathematical precision.

First of all, cognitive psychology is a research program to explain intentional behavior. Behavior patterns are explained as directed to a certain goal, governed by propositional attitudes: beliefs and desires. An action is caused by a desire to reach a goal together with a belief of the organism that the goal could be reached by producing that action, the relation between attitudes and action being rational and intentional. One thought of these relations as being matched by unconscious computational relations between symbols in the mind/brain.

In this light the process of reasoning could be looked upon – and empirically studied – as a process of problem solving: searching through a space of possible solutions. It turned out that this search process could be successfully analyzed as a series of computational operations on the organism's beliefs, resulting in a process of accepting and rejecting different beliefs.

Fodor

In 1975, Jerry Fodor's book 'The Language of Thought' (Fodor, 1975) marked a basis for the research program of cognitive psychology. The main idea was that the processes of the mind should be seen as computational processes. However, computation presupposes a representational system. A controversial thesis of Fodor claimed that every human being is born with a representational system that is basically the same for every human being. This system should be seen as a descriptive language. Within this representational system computational operations preserve properties of beliefs such as truth and reference.

Fodor put forward three empirical arguments to support this claim. The first pointed out that there is a semantic parallel between thoughts and sentences. The meaning of words can be compared with the meaning of mental concepts and sentences can be compared with thoughts. The second argument stressed the syntactic parallel between language and thought. Thoughts as well as sentences are productive and systematic. There are indefinitely many and complex types of possible sentences based on a lexicon and a syntax. The same holts for thought with a conceptual lexicon and mental rules.

The third and most important argument is the processing argument. Fodor argued that the learning of concepts is only conceivable as a process of inductive extrapolation: the formation and confirmation of hypotheses. So the learner must have a representational system that is capable of expressing the hypothesis before learning. And once concepts are learned the representational system is needed to consider and judge possibilities when it comes to a rational choice. Fodor further argues that perception is only possible if several hypotheses are considered to identify what is seen, because recognition of objects in the world is underdetermined by the raw data received by the senses.

These arguments led Fodor to the controversial conclusion that the only conceivable way of learning and using language was by already having some representational system or knowing some language: the language of thought. By further analysis of linguistic and psychological data, Fodor tried to show that the language of thought is at least as rich as any natural language. That implied that seemingly all basic concepts are hardwired in the brain. During youth we would learn to translate them into a culturally induced natural language. So, by studying language and its use empirically, we could find out the structure and operations of the language of thought.

Summary

To summarize Fodor's thesis: a part of human behavior is considered as intentional. Cognitive psychology provides an explanation of intentional behavior as governed by propositional attitudes, *i.e.* beliefs and desires. Part of the beliefs are reached by the process of reasoning. Reasoning is explained as a computational process in a representational system. All human beings are born with the same basic representational system: the language of thought. Learning is a process of forming hypotheses and confirmation within that representational system. Rationality in thinking and behavior can be seen as problem solving: a heuristically guided search through a space of possible solutions.

Today, in cognitive science, Fodor's overall thesis is criticized as well as cherished. While knowing that Fodor's ideas are open to and under criticism that I have not mentioned, I still think that they show that theories in cognitive science have implications for theories in the philosophy of science. In the succeeding section I will show that the framework of cognitive psychology is rich enough to incorporate issues of philosophy of science, as set out in section 2.2. In section 2.5, I will argue that the framework is even richer.

2.4 Interaction

In section 2.2, I interpreted the theories of some important philosophers of science as being primarily concerned with justification of scientific theories and activities. In later discussions in philosophy of science the actual practice and behavior of scientists is considered as well. It is sociologically and historically studied what theories were accepted and developed and for what reasons. This empirical work resulted in data that were not consistent with the earlier logical notions of rationality.

Later philosophers tried to show how theories about rationality could still be consistent with sociological and historical data. As a consequence, they put forward different ideas about how science develops and what the resulting theories represent, if they represent anything at all, and if or how they should be justified.

Many philosophers of science who take science as a rational business take psychology as incompetent to say anything about it. Psychologically explanations of behavior should have nothing to say about how to do science rationally. But cognitive psychology not only allows rationality as an explanation of behavior, it also explicitly studies it. It even has the potential to explain notions of rationality that are normally considered the concern of modern logic.

Yet, it could be argued that what philosophy of science should contemplate about is how to reason according to modern logic, not how people actually reason. However, since Kuhn, philosophy of science cannot leave out science's practice without inviting the argument that philosophy of science, in that case, has nothing to do with real science.

In this section I will explore how or if cognitive psychology, as set out in Section 2.3, clashes with the theories of Carnap, Popper, Lakatos, Laudan and Hacking as set out in Section 2.2. I will try to show that cognitive psychology can be a worthy opponent in issues of philosophy of science.

Carnap

Fodor's cognitive psychology is maybe closest to the logical positivism of Carnap, but at the same time totally different. Fodor's internal basic representational system shares many of its logical properties with Carnap's ideal formal language, with the main difference that the latter is an artificial logical language and the former is supposed to be a phenomenon that can be empirically investigated.

The justification of the propositions of scientific theories by confirmation seems to resemble the non-demonstrative learning process of concepts in Fodor's representational framework. Computational steps and their results in a cognitive process do not need a metaphysics for their explanation, just as the logical implications of scientific theories in Carnap's doctrine do not need such an explanation. And if, in an internal representational system, concepts are learned by hypothesis formation and confirmation, then the justification of them is induction, which was enough for the logical positivists.

Another important difference is that logical positivism did not take into account actual scientific practice. It is indifferent to any explanations of the behavior and practice of scientists; these were, in that time, governed by behaviorism. Thus one is led to the biggest difference: the terms in Carnap's language had to be purely observational, while the terms of the language of thought are theoretical. They provide an explanation for certain observable cognitive phenomena: language and intelligent behavior. But if Fodor's thesis about cognitive psychology is assumed then a form of logical positivism could be compatible with it, because an internal basic representational system implies the possibility of an ideal formal language that can provide certainty within it.

Popper

Being compatible with a part of logical positivism does not, for cognitive psychology, imply being totally incompatible with the critical rationalism of Popper. The realization that induction does not guarantee absolute certainty is a logical truth that can be justified in the frame of cognitive psychology.

To start with, there is a difference between the learning of a natural language and the justification of scientific hypotheses within a language. Concepts are a kind of theories about what to expect about instances of that concept. But concepts in natural language are not learned that strictly. Natural language is full of 'falsified' concepts that are entertained anyway. For example, a penguin is a falsification of the concept bird in English, because you expect a bird to fly. But doing science is another process: it can be seen as striving to justified knowledge within the conceptual frame of a language.

Again, first you need a language to state your hypotheses in, before you can test them. Popper argues that one should accept only singular statements, which falsify or corroborate a hypothesis, but this can only be possible within an already known language frame. The goal of developing logically justified hypotheses can only be justified within the logic of a language, and so within the language of thought. Because logic is a characteristic of the language of thought. What can be said of a science is that it develops its own language that tries to be as logically correct as possible, but again, only within a shared mental framework.

The thing that is in conflict with Popper's ideas is that the language of thought introduces the possibility for a true logic of discovery. The operations of justification by corroboration and falsification can be seen as general operations in a process of problem solving. Finding a theory or law that governs the accepted empirical singular statements about a phenomenon can be explained as solving a problem within the frame of a language.

It still leaves room for sudden insights. But because they have to be justified within the terms of the (scientific) language, sudden insights can be seen as suddenly finding a solution within that linguistic frame. In studies of in cognitive psychology it is found out that the process of finding a solution to a problem is not a mere process of trial and error. It can be seen as a heuristically guided search through possible solutions.

Yet, while many scientific discoveries occur within a frame work of a given scientific language, many revolutionary discoveries are accompanied by a change in the conceptual framework of a scientific language. This will only clash with Fodor's thesis if one accepts that every basic term in the language of thought is also a basic term in the frame of a scientific language. However one could argue that one should understand the basic terms, that according to Fodor are needed to explain the whole process of learning a language, are present on a different level of abstraction. In a similar case, the psychologist David Marr argued that we need to assume that the projection of cylinder forms on perceptual data is hard wired in the brain to understand the process of object recognition (Marr, 1982).

Kuhn

One can summarize Kuhn's view of science and his reaction to Popper by stating that there can be no logic but only psychology of discovery. But it is incorrect to conclude that therefore science is not rational or can not be understood as a rational enterprise. With Fodor's thesis about cognitive psychology one can provide an explanation of rationality in science.

Entertaining the concept of the language of thought implies a common logical basis for all possible paradigms of science because theory development and justification is done by human beings who share a common basic conceptual system. Thus, supported by empirical data of cognitive psychology, the acceptance of the language of thought makes incommensurability and the implied relativism unjustified concepts.

Kuhn is still right in rejecting an early form of falsificationism for explaining science, but wrong in rejecting the possibility of a justified form of rationality by just asserting that it is a matter of psychology, because then he clearly underestimates the reach of cognitive psychology. It explains the behavior of individuals as well as their behavior in the context of a paradigm without resorting to social forces only. Of course, it still remains a point of discussion whether it provides a proper explanation.

Lakatos

As an heir of Popper, Lakatos shares his objections against psychology, but again, also without recognizing that rationality can be justified within cognitive psychology. So his refinement of Popper's falsificationism, by allowing an irrefutable theoretical core and letting auxiliary hypotheses take all the refuting blows, can be comprehended in a cognitive psychological frame just as well.

What is incompatible with this frame, is that Lakatos' refinement might allow changing protecting hypotheses forever. But, if the core assumption of a research program is incorrect in respect to the world and the language of thought, changing auxiliary hypotheses would eventually turn out to be empirically unjustified or would otherwise lead to a change in the relation between the language of thought and the scientific language in which the theory is put.

The correctness of a theoretical core can result in unmasking presupposed hypotheses. But if protecting hypotheses would be continuously changed to save the core then all hypotheses eventually lose their meaning (and thus the possibility of comprehension within the language they are put in is lost as well), because they lose their initial relation with the language of thought. Without that relation the theory would not make any sense for anyone knowing the initial scientific language.

Laudan

As for the comparison with Laudan's ideas, cognitive psychology incorporates an explanation of the notion and usefulness of problem solving in science. It entertains these as a basic feature of human cognition that can be rationally guided.

An explanation of why one theory solves some problems better than another can be that, given a scientific language, the one is more truth-like then the other: it is a better possible solution then the other within the space of all possible solutions, within a learned instantiation of the representational frame of the language of thought.

So, successes or progression with problem solving can be explained by presupposing a cognitive process that is, for having success, governed by truth in a representational system.

Hacking

When we see Hacking as accepting only parts of theories, their theoretical terms when they can be manipulated during experiments, we see that this is again a point of view that can be incorporated in a psychological frame, as all other philosophies reviewed so far. One just has to regard the possibility of intervention as a form of rational justification of the reference of terms within a language. From there to, how language relates to the language of thought, it is the same story as above.

Summary

I put forward Table 2.1 as a synopsis of the stances regarding issues in the philosophy of science with Fodor as a participant. I regard the different philosophers as having a philosophical and/or empirical theory about science consisting of: a theory about the scientific practice or behavior of scientists; a theory about what theories represent; a theory about when and how theories develop; and finally a theory why scientific theories are justified or accepted.

	Practice	Representation	Development	Justification
Carnap	-	Empiricism	Confirmation	Induction
Popper	-	Realism	Corroboration	No falsification
Kuhn	Paradigms	Relativism	Normal/revolution	Puzzle solving
Lakatos	Research program	Realism	Progression	Empirical content
Laudan	Research tradition	Pragmatism	Progression	Problem solving
Hacking	-	Entity realism	-	Intervention
Fodor	Prop. attitudes	Internal realism	Confirmation	Problem solving

Table 2.1: Different views on science

Carnap and Popper did not consider scientific practice because they mainly pursued a normative philosophy of science. Kuhn introduced paradigms into the picture and showed how the ideas and behavior of scientists depends sociologically on the scientific paradigm they work in and on. Lakatos adjusted Kuhn's paradigms and explicated the organization of scientists around the theoretical core of a research program. What a scientist accepts depends on the program he is working in. Laudan further extended research programs to research traditions and included, among other things, the conceptual world view of scientists that also determined their adherence and work in a certain tradition. Hacking speaks about scientific practice but he does not pretend to explain it. If we regard Fodor then we should not look only at a sociological level: we can also explain the behavior of the individual scientists on a psychological level, in terms of their propositional attitudes.

What scientists do, should do, or can do is dependent on what their theories represent. Carnap and Popper both thought that theories represent the world with the main difference that Carnap's logical positivism, also called empirism, only allowed theories that represented, or could be redescribed to represent, things that can be observed. Popper's realism had less problems with theories representing unobservables. The relativism of Kuhn on the other hand does not regard theoretical terms or even observational terms as representing anything in the world, because when a theory is developed further the meaning of its terms change as well. Lakatos is as much a realist as Popper was. Laudan, on the other hand, regards theories as useful but does not allow representation. Hacking does allow the reality of 'theoretical' entities if it can be shown that they cause something in experiments. Concerning representation, Fodor allows a realism that is justified with the internal representational system of the language of thought. The terms of theories represent the world in respect to human perception, the language of thought and the particular language a scientist employs. The language of thought thesis incorporates the idea that perception is theory laden, it deals with underdetermination and undermines incommensurability.

A further important philosophical problem is the question when, how and if science develops. According to Carnap we have learned something about the world if a new hypothesis is confirmed. Popper speaks of the corroboration of hypothesis which stood up to critical tests. In both views progression is the result. But when Kuhn looked at history he only saw normal science and revolutionary leaps between incommensurable paradigms in the development of science. Lakatos, however, explicated that when a research program develops the empirical content of the theories should increase. For Laudan there is just progression when a tradition does not run into unsolvable problems. As far as I know, Hacking did not propose his own theory about how science develops. For Fodor, a person learns truths about the world, as well as a natural language, when hypotheses are confirmed in his internal representational system.

The last philosophical problem taken into account concerns the justification of scientific theories. Carnap justifies confirmed hypotheses by induction. But Popper is only willing to pursue a theory when it is not falsified. For Kuhn a theory is still worthy if puzzles can be solved with it during a period of normal science. Lakatos admits theories as long as the research program keeps on increasing its empirical content and does not degenerate. For Laudan problem solving is the goal of science, not truth. Hacking justifies the acceptance of theoretical entities when we can explain the result of our intervention with nature in terms of them. And finally, Fodor's cognitive psychology admits the acceptance of theories partly by induction but mostly by problem solving. The descriptive and explanatory nature of the language of thought does however allow a normative bent: human problem solving can be analyzed and improved.

This section tried to show that issues and research problems of cognitive psychology can be considered as part of the problems and issues of the philosophy of science. The ideas about the issues clash, but in the same way as the theories within philosophy of science clash with each other. In the next section I will regard how philosophy of science can fall within the frame of cognitive psychology instead of the other way around.

2.5 Integration

Up till now, I argued for a place for cognitive psychology within the philosophy of science. But you can also look at issues of philosophy of science as constituting a part of the issues of cognitive psychology.

As we saw in Section 2.3, Fodor's cognitive psychology is concerned with the explanation of intentional behavior and cognitive processes that result into language and rationality. This is accompanied by an explanation of understanding and comprehension of symbols and the world. Consequently, within the frame of cognitive psychology, scientific theories, as all other symbols, should be processed by a person's mind to have any meaning. Their reference is determined by the person's representational system. That makes truth a feature of the cognitive processes of the mind. Hence, cognitive psychology can in fact be seen as a scientific epistemology.

Theories in the philosophy of science, in that way, can be interpreted as theories about cognitive processes within a representational system that can be empirically investigated within the frame of cognitive psychology. The study of the relations between theories on different levels of explanation would be a study of the processes of thought within the mind and of the mind's representational system. In this way, epistemology can be seen as a science about how human beings know the world, and can learn to know it better. The foundation of knowledge would not lie exclusively in perception of the world, neither would it lie in language, it would lie in how human beings can know about and act in the world on the basis of their representational system: it would lie in the language of thought.

The framework of cognitive psychology is even richer than its capability to explain theory justification, it can explain theory discovery as well. It gives the possibility to study justification and discovery within the same framework. One way of doing so is understanding discovery as the result of an heuristically guided search through a space of possible solutions of a given problem. That problem could, for instance, be: what formulas explain the given empirical data. Investigating the process of discovery would then be investigating how scientists learn heuristics that can find solutions for a problem in a given representational system. In that way it can be seen that justification is also an operation in the process of search on the level of discovery and not just a judgement after discovery.

It is not the case that cognitive psychology does not allow “the spark of brilliance” or any other notion that is related with serendipitous discovery and creativity. There can be more ways of finding a solution in a space of possible solutions than only through a methodological search that is heuristically guided. But the hypothesis of the language of thought implies that: if a solution to a problem can be found serendipitously in a given finite problem space then it can also be found by method.

Finally, the program of cognitive psychology was originally set up to explain individual behavior of human beings, so it may allow us to give an adequate reconstruction of the behavior and ideas of scientists. It is very likely to meet the challenge to justify historical and social data, regardless whether the goal of a scientist is power or truth.

Fodor’s thesis in cognitive psychology provides a theoretical frame for processes of cognition that can explain features of language, thought, and behavior. Those processes are seen as computations in an internal basic representational system. From this viewpoint, theory discovery, development, and justification next as well as intentional behavior of scientists are all governed by computational processes in a representational system which can be empirically investigated. So, cognitive psychology implies a stance within the philosophy of science because the assumptions of the philosophy of science are a part of the assumptions of cognitive psychology.

Hence, we can consider issues of the philosophy of science as part of the issues and research problems of cognitive psychology. Theories of science can be interpreted as theories about certain cognitive processes and their desired results. In the next section I will close this chapter with a general conclusion.

2.6 Conclusion

Should someone accept the theories of a science because he accepts the ideas of a philosophy of science that justify that science? Or should one accept a philosophy of science because it is justified by the science that is accepted?

I tried to make clear in this chapter that, if you accept some idea in the philosophy of science, you implicitly accept some psychology or philosophy of mind, and if you accept some psychology or philosophy of mind you also accept some philosophy of science. They are both about human knowledge and reasoning. If you state how a theory can be justified, you presuppose how a human being can represent theories as well as where they are about. If you state how human beings can have knowledge, and how it influences their behavior, you presuppose how human beings can justify their knowledge. For making this claim, I used the assumptions about cognitive psychology by Fodor and related them to issues in the philosophies of Carnap, Popper, Kuhn, Lakatos, Laudan and Hacking, showing that Fodor and these philosophers of science share a number of issues and assumptions.

What is important for this claim is that it does not matter whether you accept Fodor’s cognitive psychology or not, it holds for every scientific theory about language, thought, behavior, and the brain. If you do not accept a non-physical theory of psychology, but only consider neuro(physio)logical information processing, you still presuppose some account of justification of theories. But if you regard the processes of the mind and brain as computational then you should see a scientific theory as a computational recipe in a representational system, *i.e.* as a kind of computer program. Looking to theories from that perspective opens up a whole world of new ways to study and understand science with the aid of theories about representation, computation, learning, rationality and behavior within cognitive science.

As a conclusion, I will repeat the claim I argued for in this chapter: cognitive science in general can make sensible contributions to debates, ideas and developments in the philosophy of science because accepting a theory about science is accepting a theory about the mind/brain and vice versa, philosophically as well as scientifically. How psychology can contribute to the debate about the rationality of science is a main topic of the rest of this thesis.

Postlude

Philosophy of Science realized what his young nephew Cognitive Science had in stock. Now they both had to convince their own families about their combined potential. The best way to achieve that was getting to work together and let it be shown…

* *

Chapter 3

Neuropharmacology

3.1 Introduction

What is the rational use of theory and experiment in drug research for Parkinson's disease? In this Chapter I discuss this problem from a bird's eye perspective, providing an introduction to the more detailed analysis of discovery in neuropharmacology in Part III of this thesis.

The approach of this thesis is that the best way to understand the process of discovery in empirical science is to see it at work. This opinion is endorsed by both psychologists, studying how people actually make discoveries in scientific practice (*e.g.* Dunbar, 1995), and computer scientists, who want to make programs that aid discovery, no matter how people actually make discoveries (*e.g.* Valdés-pérez, 1998).

I took a similar approach, in conducting my case study of drug research for Parkinson's disease at the Groningen University Center for Pharmacy. It turned out that fundamental research into the biological mechanisms of the brain and new drug experiments go hand in hand in the search for new drugs. Theories and models of biochemical and neurophysiological mechanisms guide the search for a new drug and drug treatment, and newly designed highly selective drugs are used to empirically test those models and further explore those mechanisms in the laboratories.

This chapter globally surveys my analysis of the reasoning involved in using theoretical diagram models in neuropharmaceutical research. These describe relations between variables of a biological system. The use of such diagram models has some limitations in practice, due to their complexity. A formal way to understand these models is to represent a model as a qualitative differential equation. An explication of the reasoning task can help to understand the search for drugs led by suggestions originating from such models, and possibly aid that task by computational techniques.

The next section briefly describes the field of neuropharmacology and the case of Parkinson's disease. In section 3.3, I will compare the reasoning in the search for an explanation with reasoning in the search of a drug treatment. In section 3.4 I outline a method of making predictions from knowledge in neurobiology by using qualitative differential equations. Section 3.5 defines and discusses the process of rational drug discovery. This chapter ends with some general conclusions.

3.2 Description

In this section I globally describe the field of neuropharmacology. One aim of drug research in neuropharmacology is to find a way to intervene in neurophysiological and neurochemical processes such that pathological properties or symptoms are suppressed, or desired properties are induced, (Vos 1991). Those unwanted properties are determined and discovered in numerous ways. The history of pharmacology and medicine is rich with serendipitous cases where a patient with a particular disease comes into contact with a compound that enhances his condition, hence providing a clue about the disease mechanism. A systematic study involves comparison of properties of pathological processes of patients with those of control subjects. In some cases, such as in Parkinson's disease, a cause of disease symptoms can be traced back to different concentrations of a single neurotransmitter compound.

Neural disorders have their origin in shifts in delicate balances of neurochemicals, which can be caused by *e.g.* cell damage or degeneration. The plasticity of the brain is large enough to restore imbalances, *e.g.* by increasing the sensitivity for a particular neurotransmitter. But when it fails, *e.g.* when a substance is depleted almost completely as in the case of Parkinson's disease, a severe neurological disorder results.

The aim of a therapeutic strategy is to find or design chemicals that selectively influence neurotransmission. The goal is to restore balances by administering those chemicals, to nudge derailed processes back on the track. This kind of research has a top-down and bottom-up strategy. In the latter case, one tries to discover and understand structures and processes in the brain by influencing them selectively, and seeing what happens. This is done both locally and globally. How does a new drug influence local neurological processes, and how does it influence behavior? In the top-down case, one uses all knowledge available about the pathology of a disease to discover new therapeutic targets, leading to a so-called *drug lead*. This is a description of the functional properties a potential drug should have to influence that target. In practice, top-down and bottom-up go often hand-in-hand.

Using knowledge to build models of neurochemical structures and processes to guide drug research is dubbed rational drug design. Computational models of complex receptor structures are made to infer what chemicals might interact with them. Yet, in contrast to such rational methods, currently a very successful strategy is to generate chemicals massively and to test them *in vitro* on their potency for influencing receptors. This strategy will end up with a nice set of chemicals to influence the biological machinery in a highly selective way. On the other hand, it is not always obvious how to employ those chemical tools optimally. For example, it may turn out that a particular combination of drugs is needed to properly influence several mechanisms involved in a disease. This can be discovered by first rationally understanding those mechanisms.

Hence, fundamental research into the workings of the mechanisms of the brain is also pursued in neuropharmacology. One research tool employed is building models of neurochemical and neurophysiological processes that aim to fit data acquired by lab-studies on animal models. This is conducted in the Pharmacy Department of the Groningen University by employing electrophysiological methods and microdialysis to track nerve signals.

A nerve propagates a signal by conducting an electric pulse called an action potential. This signal initiates the release of transmitter chemicals at the terminals of the cell, that affect receptors of nearby nerve cells that may further propagate a signal. Placing an electrode in the brain can monitor the electrical activity. The release of transmitter is measurable by means of a microdialysis probe. This probe can also be used to release chemicals locally and measure the effect *in vivo*. At the Pharmacy Department of the Groningen University the function of neurophysiological pathways is studied by using these two techniques.

Specific studies of the functional relation between several variables together contribute to understanding the function of a brain area, or cell groups called nuclei. To describe these neural circuits, box and arrow models are drawn showing positive and negative influence relations (Timmerman, 1992). These models are further tested for their correctness and used to explain and predict the functioning of the system. Newly developed drug compounds play a bootstrap role in this research: they are used to revise and refine the model and experiments conducted, while on its turn the model is used to understand their effect. A drug that works very selectively for one particular type of pathway can be used to further explore the function of that pathway. The acquired data may then serve to refine the model, so that the effects of the new drug can be explained and predicted.

A group of subcortical nuclei called the basal ganglia are studied in Groningen (Timmerman *et al.*, 1998). These nuclei play an important role in the control of voluntary behavior. In the case of Parkinson's disease a part of them, called the *substantia nigra pars compacta* (SNC), decays due to an unknown cause. The SNC is a supplier of an important neurotransmitter called dopamine, which is postulated to serve a modulating function. It is thought to maintain a delicate balance in influencing signals from the cortex. To understand this balance a schematic model is used to represent neural activity in the basal ganglia in Parkinson's disease, see Figure 3.1.

Figure 3.1 presents a schematic representation of neural activity in the basal ganglia in Parkinson's disease, as postulated in studies by Timmerman (1992, p. 18). An arrow in the diagram is a neural pathway, consisting of a bundle of individual nerve cells. A box is a nucleus, or clustering of nerve cells. Increased inhibition induced by receptors sensitive to the transmitter GABA of the external segment of the *globus pallidus* (GPe) leads *e.g.* to disinhibition of the subthalamic nucleus (STN). In turn, this provides increased excitatory drive to the internal segment of the globus pallidus (GPi) and *substantia nigra reticulata* (SNR), therefore leading to increased thalamic inhibition. This is reinforced by reduced inhibitory input to the SNR/GPi. These effects are postulated to result in a strong inhibition of brainstem neurons. D1 and D2 are two different types of receptors, postulated to react excitatory and inhibitory, respectively, to dopamine (DA). (This model is explained in more detail in part III)

In the model dopamine has a dual function. It enforces the direct path from the striatum to the SNR/GPi while it inhibits the indirect path, via the GPe and STN. This balance maintains an inhibition of both the brainstem and the thalamus. Yet when dopamine is nearly depleted, the balance becomes disrupted, resulting in a strong increase of the activation of an area called the SNR/GPi, see Figure 3.1. This hyper-activation causes strong inhibition of brainstem neurons and is correlated with some of the major symptoms of Parkinson's disease.

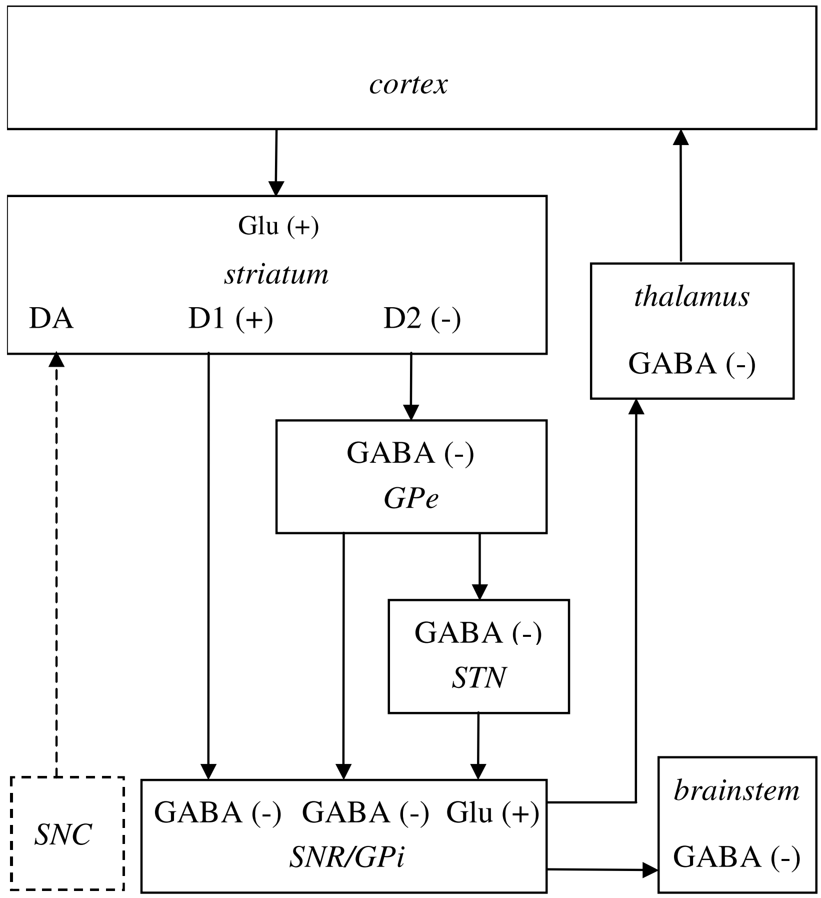

Figure 3.1: Diagram model of the basal ganglia

Most of the traditional research on Parkinson's disease is focused on restoring levels of dopamine. This compound cannot be administered as a drug that can be swallowed because it does not pass the so-called blood-brain barrier. Yet it was discovered that L-dopa, which metabolizes in the brain to dopamine, can pass this barrier. Administering doses of L-dopa regularly is to date the most successful therapy to deal with Parkinson symptoms.

Yet administering L-dopa also causes dopamine levels in other parts of the body to increase. This higher concentration of dopamine in the blood causes nausea as a side effect due to stimulation of dopamine-receptors elsewhere in the body. And after three to five years of use the therapeutic effect wears off drastically. Further research investigates the use of highly selective dopamine receptor agonists, compounds that interact only with particular dopamine receptors. The dopamine receptors on the direct route from the striatum to the SNR/GPi were discovered to be mainly of another type (D1) than that of the indirect route (D2) via the GPe. Both receptors can be stimulated by dopamine, but with different effects. D1-receptor stimulation with dopamine has an excitating effect on a cell, while stimulation of the D2-receptor with dopamine inhibites the cell. Clinical studies are conducted to investigate the therapeutic effects of using different compounds that differ in selectivity to both the D1 and D2-receptor. These studies show that using only a selective D1-agonist, a compound that stimulates D1 but not D2-receptors, is not successful.

The model in Figure 3.1 is used to understand the effect of selective compounds. However, in the literature opinions about these kinds of models are rather diverse. Some people use them to understand and theorize about physiological phenomena extensively, while others are wary of using them because they are too simple, not respecting the subtlety of the data, and therefore not realistic. In a recent article in the movement disorder literature it is said:

> "On the one hand, efficient models have to be simple, but simple models can provide only part of the reality and are thus bound to be wrong (for example, current basal ganglia model) ... On the other hand, an elaborated model that would embody all the complexities of a given reality ... is doomed to be useless" (Parent and Cicchetti, 1998)

The practical problem of the diagram model is that it is informally represented. Its consequences are inferred by tracking the boxes and arrows. The general basal ganglia model is already fairly elaborate. A more realistic picture would have to be substantially larger, including more transmitters, peptides, small interactions and feedback loops. Including these would cloud the overview, drowning it in the complexity of all the consequences of the model.

The following sections generally describe a part of the reasoning involved with such models, introducing the use of qualitative differential equations to represent them. These allow for systematic and computational exploration of their consequences and have the potential to aid with both the understanding and the testing of the models, but also to explore them for new drug lead suggestions. But first we will look at the kind of reasoning that is involved.

3.3 Explanation

In the literature on scientific discovery, a lot of attention is paid to understanding and explicating the process of explaining surprising or anomalous observations. The generation of potential explanations is often dubbed *abduction* after the work of C. S. Peirce, whereas their evaluation is known as *inference to the best explanation*. The starting point in those analyses is in most cases a new phenomenon or observation that comes as a surprise, because it cannot be explained by current knowledge, or because a contradictory outcome was predicted. From then on, new explanations are sought, evaluated, and incorporated in the known theories and background (Th. Kuipers, 1999). How an anomalous or surprising observation comes about is often the result of casual observation, serendipity, or devised laboratory experiments that aim to test explanations on their correctness.

In pharmacology, the research aims have a strong pragmatic component. The goal in rational drug design research is to understand a particular biological structure or mechanism and to use this knowledge to devise chemicals to influence it. A research problem aiming at a new drug treatment for a particular disease starts with phenomena or symptoms that do occur but that we do not want to happen, or with properties or symptoms that do not occur, but that we *do* want to see. Now the goal is to find a drug inducible condition that causes the wanted properties to occur and the unwanted symptoms to disappear. Superficially this reasoning task does not differ in structure from abduction and inference to the best explanation. Only the status of the initial condition and the observation is different. In the explanation task, the observation to be explained occurs, needing an as yet unknown initial condition or theory to explain it. In the other case, the wanted property does not occur, needing an as yet absent initial condition that can cause it to occur.

The search involved is structurally similar to that of abduction and inference to the best explanation, but it has a different goal. Instead of finding a simple explanation of an observed effect, the task is to infer a simple (drug) intervention that causes a desired effect. So we could call this reasoning task: *inference to the best intervention*. Note that this task differs from diagnostic reasoning. Inferring what causes a disease symptom is not the same as to infer how to remedy it. That may often be as simple as removing the found cause, *e.g.* by killing a germ. But, as the case of Parkinson's disease shows, that is not always possible.

3.4 Prediction

To understand inference to the best intervention based on the schematic diagrams about the dynamics of the brain we employ the formalisms of qualitative reasoning to deduce predictions from those diagrams (*cf.* B. Kuipers, 1994). In qualitative reasoning research, the structure of a dynamical system is described by a qualitative differential equation (QDE), that defines the relations between the variables of the system. The exact nature of a relation may not be known, as is the case in many investigated relations in the model in Figure 3.1. Yet it may be known what the *sign* of the relation is. It may be known that a function describing the relation between two or more variables, that change in time, belongs to the class of monotonically increasing (M^+), or decreasing (M^-) functions.

Furthermore, any variable can be ascribed a qualitative landmark value such as *high*, *low*, or *normal*, and a direction of change over time: *increasing*, *steady*, or *decreasing*. Several variables can influence one other variables such that the differentials of all variables together determine the resulting value. There is a calculus defined to determine these values. For example, if the value of variable p_1 is a differential function over time of p_2 plus p_3 and the function belongs to the class of monotonically increasing functions, then the value of p_1 will increase if both p_2 and p_3 increase, but remains unknown if p_2 increases and p_3 decreases. The lack of knowledge in the last case is a necessary consequence of the qualitative and incomplete character of a QDE.

A qualitative state of a system described by a QDE is an attribution of variable values to all variables of the system, consistent with the constraints in the QDE. Given a QDE and a set of known initial variable values, a set of all consistent system states can be deduced, together with their possible transitions. When a calculated value is unknown, all possible states are included in the set. This set is complete, but is proved to be not always correct since spurious states may be included as well.

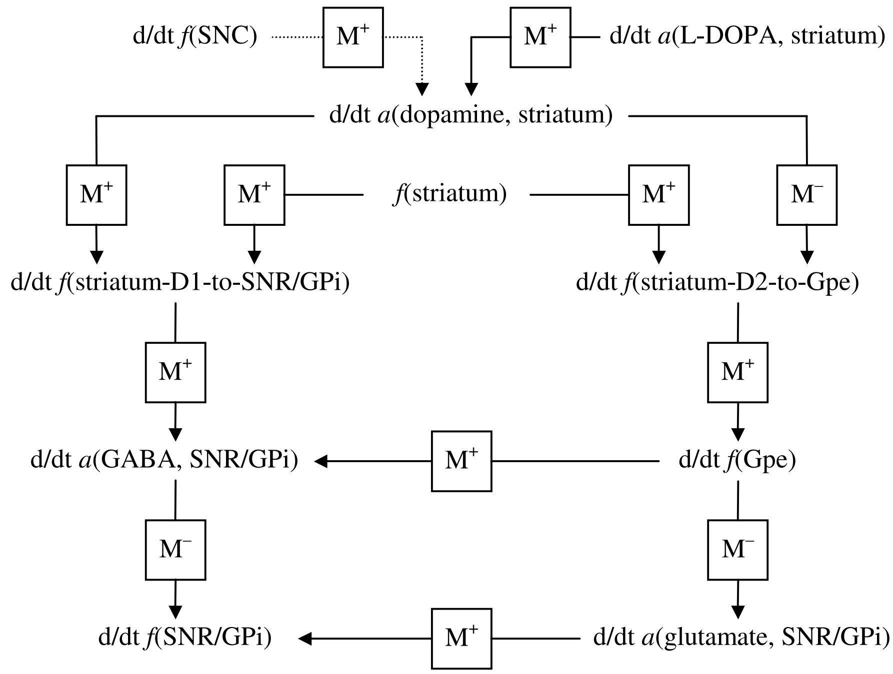

Figure 3.2: QDE fragment of the basal ganglia

Figure 3.2 displays a QDE fragment including a part of the basal ganglia model in Figure 1, and the metabolism of dopamine. It relates variables such as the firing rate (f) of nuclei and neural pathways, and amounts (a) of neurotransmitters in nuclei. For example, the increase of the firing rate of the SNC causes an increase in the amount of dopamine in the striatum, while this latter increase causes a decrease in activation of the neural pathway that signals to the GPe, etc.

3.5 Intervention

In medical practice, a disease is characterized by a profile, which is a set of characteristics with certain qualitative values. Given a profile, it is a goal in neuropharmacology to discover a drug lead, which is a set of wished-for functional drug characteristics (Vos, 1991). This search can be based on qualitative knowledge if the profiles include comparative values of variables of a normal and a pathological state of a system. This is the case when values of variables are known to be higher or lower in a pathological condition, compared to controls.

The search goal is to find those variables by which one can intervene in the profile in such a way that the pathological values of the variables associated with a disease are reversed. The goal set is defined to consist of the variables of the disease profile with an inverted direction of change, *i.e.* if a variable value is lower in the pathological profile, it is included in the goal to increase that variable value. We can now define the ideal goal of this search task: find a minimal set of variables such that a manipulation of the variable values propagates a change in direction of the values of the variables in the goal set.

However, there may not exist a set of variable influences that causes all desired changes of values of the goal set. So we have to moderate our goal to find that set of variables for which an influence causes the largest number of desired goal variable values, while minimally affecting the other variables of the system. This intuition can be explicated by an approximation criterion analogous to a criterion used in explicating design research and truth-approximation, (T. Kuipers, 1999; Van den Bosch, 1997, 1998, see part III).

The defined task can now be carried out as a search in a solution space of conceptually possible interventions. We start with a QDE model and known initial values of its variables. A goal of desired variable values is set. Reasoning backward from the goal values one can explore possible manipulations of the variables. The approximation criterion is used to measure the difference between the goal values and the values caused by a particular manipulation, implementing a means-end analysis.

In Parkinson's disease, the goal set includes a lower activation frequency of the SNR/GPi than in the pathological case, *cf.* Figure 3.1. A search through possible manipulations will not only find an increase of the amount of L-dopa in the striatum. It will also find that a decrease of the firing rate of the indirect pathway between the striatum and the GPe results in a decrease of the firing rate of the SNR/GPi. Administering a selective D2 agonist can cause such a decrease, with a lesser effect on other dopaminergic pathways than dopamine.

This reconstruction tells us nothing new about what to do about Parkinson's disease. Yet by making the knowledge and reasoning explicit (by describing it formally) it is possible to increase the complexity of the basal ganglia model without rendering such a model useless in the manner that was argued in the movement disorder literature. Via a computer program as a modeling tool it is still possible to keep track of, and further investigate, all the consequences of such a model.

However, because of the incompleteness of the data, numerous and possibly spurious suggestions will be made. So, drug lead suggestions can best be seen as proposals for experiments. A manipulation derived from current knowledge is an excellent basis for a new experiment design serving both a practical and epistemic goal: testing a manipulation for its therapeutic appropriateness and testing the models used to derive the manipulation for their correctness.

If a large enough domain of data is included, it also has the benefit of connecting results, in the way the ARROWSMITH program does, based on text analysis of titles in the MEDLINE-abstract database (Swanson and Smalheiser, 1997). ARROWSMITH discovers the missing link between literature that describes relations between subjects, compounds or functions A and B and literature that did the same for B and C, but in ignorance of each other. In this way the relation between magnesium deficiency and migraine was discovered, via eleven intermediate effects linking them together. In principle, inference to the best intervention can do the same, given qualitative models of results in MEDLINE. Initiatives to collect results in biology in qualitative formalisms on a grand scale are already undertaken; see, for instance, the EcoCyc and MetaCyc projects on the web by Karp and Riley (1993).

3.6 Conclusion

The rational use of neurophysiological models can be modeled as goal directed reasoning about qualitative differential equations. Applying effective search techniques to such models could potentially aid drug lead discovery for complex biological systems with a large set of variables and constraints. However, this is a claim only warranted by theoretical considerations. Whether novel results can thus be produced still has to be seen, because there are problems as well. When a large-scale QDE model is compiled it can be severely inconsistent because the empirical results are not always mutually consistent. Yet by using the best intervention suggestions to devise new experiments, qualitative reasoning about neurophysiological models as part of a computer supported discovery system could still aid in using, understanding and testing models about larger biological systems.

This also concludes the introduction part of this thesis. Part II will go further into rationality in discovery in more detail, while Part III will, in detail, further address discovery in neuropharmacology.

* * *

Part II Discovery

What is the rational use of theory and experiment in the process of scientific discovery, in theory? In this part I discuss three different approaches to the study of the rational use of theory and experiment in the process of scientific discovery. I start with a discussion of the study of **logic** (Chapter 4). Then I discuss an account that stems from the psychological study of **cognition** (Chapter 5). I finish this part with the discussion of a model of discovery that is grounded in the study of **computation** (Chapter 6).

Chapter 4

Logic

4.1 Introduction

In this thesis we set the general problem: what is rationality in scientific discovery? This question receives attention from several academic disciplines. Traditional philosophers of science are usually interested in what scientific discovery ought to be, and how reasoning in that process can be valid or justified. Empirical scientists are usually more interested in describing rationality in scientific discovery as a social or psychological phenomenon, to be studied empirically.

In this chapter we will address a normative approach that stems from studies in logic. In the next chapter we will address a psychological theory about the rationality of reasoning and problem solving. This part will end with a chapter on a general computational model of discovery. In discussing all models I will look for answers to the specific questions from section 1.3, *i.e.* those about: (1) the structure of a theory, (2) the process of scientific reasoning and (3) the route between theory and experiment.

In this chapter we start with a discussion of logic, the traditional study of valid reasoning. The question is: what is the rational use of theory and experiment in the process of scientific discovery, as proposed in the study of logic? We start by asking: what is a scientific theory and what is scientific reasoning?

To address these questions I discuss an illustrated example of explanation. In an episode of the life and times of cartoon character Calvin and his tiger Hobbes he watches a sunset with his father, see Figure 4.1. His father explains the setting of the sun to Calvin. Now, why would we not accept his explanation as scientific? Is this because his hypothesis is not scientific? Is this because his reasoning is not valid? Let us look at the validity of his inferences from the perspective of logical argumentation theory, and reconstruct his reasoning.

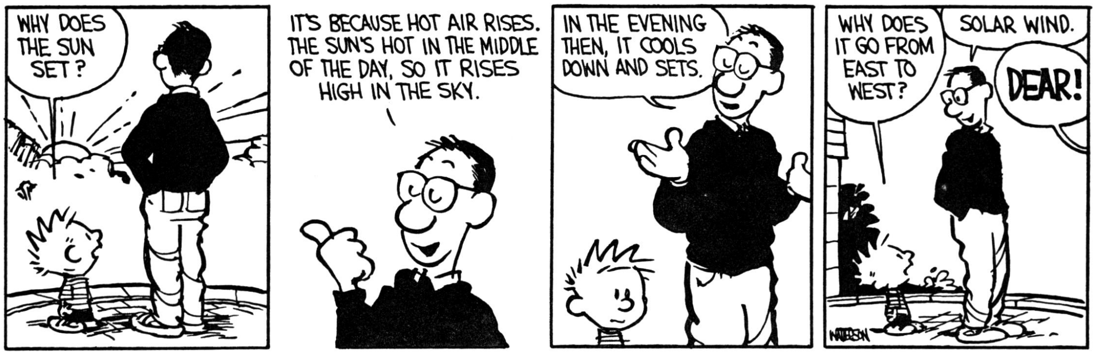

Figure 4.1: Calvin and Hobbes © 1988 Bill Watterson. Reprinted by permission of Universal Press Syndicate.

4.2 Deduction

Calvin's question is: why does the sun set? This question asks for an explanation of his observation. He wants to know what causes the sun to set. If Calvin accepts only a logically valid answer, he can only accept as explanation a deduction of his observation from what is known. Let us examine his inferences one at a time and comment on their validity.

In modern logic the validity of an inference is independent of the truth of the premises. Yet when an inference kind is valid the conclusion is true when the premises are true. To represent kinds of inference schemes in the discussion I will use a two or three letter abbreviation (TLA) that is *italicized* if it represents a logically valid inference. In an inference scheme I will mark a proposition with a star (*) to indicate that we do not know whether that proposition is true.

Calvin's father manages to infer his answer in several possibly implicit steps. First he presupposes two propositional premises as initial assumptions which Calvin should accept off-hand, without further argumentation:

P_1 Hot air rises	Hot(air) $\Rightarrow$ Rises(air)
P_2 In the middle of the day the sun is hot	Hot(sun)

Presumably he further assumes that the air is hot, and that the sun causes it:

P_3 If the sun is hot then the air is hot	Hot(sun) $\Rightarrow$ Hot(air)

These premises seem unproblematic. Based on them he can validly infer by *modus ponens* (*MP*) that the air is hot:

P_2 In the middle of the day the sun is hot	Hot(sun)	
P_3 If the sun is hot then the air is hot	Hot(sun) $\Rightarrow$ Hot(air)	
		MP
P_4 In the middle of the day the air is hot.	Hot(air)	

By transitivity (*TRN*) he can infer validly that if the sun is hot the air rises:

P_1 Hot air rises	Hot(air) $\Rightarrow$ Rises(air)	
P_3 If the sun is hot then the air is hot	Hot(sun) $\Rightarrow$ Hot(air)	
		TRN
P_5 If the sun is hot then the air rises.	Hot(sun) $\Rightarrow$ Rises(air)	

From the premises he then infers in two steps why the sun rises. To be explained first is the observation:

P_6 In the middle of the day the sun rises	Rises(sun)

This should be a conclusion from our premises and valid intermediate conclusions. For the first step three different inferences are possible. The first could be:

P_5 If the sun is hot then the air rises.	Hot(sun) $\Rightarrow$ Rises(air)	
		GEN
P_7 If the sun is hot, anything rises *	for all x Hot(sun) $\Rightarrow$ Rises(x) *	

Logically this is a fallacy, a hasty generalization (GEN) called a *secundum quid.* Seeing one type of object with a property does not imply that all have the same property. So this inference is invalid. However, we can not say that, logically, his conclusion is false either. The inferred proposition could well be true, but its truth does not follow deductively from the truth of the premise.

Alternatively, Calvin's father could have assumed that the sun is part of the air and a property of air is also a property of the sun. Since hot air rises, a hot sun rises as well:

P_1 Hot air rises	Hot(air) $\Rightarrow$ Rises(air)	
P_8 The sun is part of the air	sun part of air	
		DVS
P_9 If the air is hot the sun rises *	Hot(air) $\Rightarrow$ Rises(sun) *	

This is known as a *fallacy of division* (DVS), where a property of the whole is also ascribed to a part. All the parts together could well not have the same property as the whole (*e.g.* the parts are light, but the whole is heavy). Yet again, it is also possible that the whole does have the same property as the parts, and vice versa (*e.g.* the whole is light, therefore each part is light).

A third possible interpretation of the explanation of Calvin's father is a *causal argumentation* (CAU):

P_4 In the middle of the day the air is hot.	Hot(air)	
P_6 In the middle of the day the sun rises	Rises(sun)	
		CAU
P_9 If the air is hot the sun rises *	Hot(air) $\Rightarrow$ Rises(sun) *	

In this case a cause-effect relation is inferred from the mere observation that two events take place together. The air is hot and the sun rises, hence if air is hot the sun will rise. This is again a logical fallacy. The causal relation could just as well be the other way around, or not existent. The occurrence of events one after another could just as well be an incident. This fallacy is called *post hoc ergo propter hoc*.

We now saw three ways to infer P_9 in a first step. To explain P_6, the rising of the sun, he further infers in the second step:

P_9 If the air is hot the sun rises *	Hot(air) $\Rightarrow$ Rises(sun) *	
P_4 In the middle of the day the air is hot.	Hot(air)	
		AA
P_6 In the middle of the day the sun rises *	Rises(sun) *	

In this inference the second premise affirms the antecedent (*AA*) of the first premise. This inference is called *modus ponens*. It is a valid inference that guarantees the truth of the conclusion if the premises are both true. But in this case the conclusion may be false because the first premise may not be true. So P_6 follows validly from P_4 and P_9, but not from our initial premises P_1 to P_4, because P_9 does not follow from them.

But Calvin's question was why the sun *sets*. To explain this, his father first implies in a third step that when the sun is not rising the air is also not hot.

P_9 If the air is hot the sun rises *	Hot(air) $\Rightarrow$ Rises(sun) *	
P_{10} In the evening the sun sets.	not Rises(sun)	
		DC
P_{11} In the evening the air cools down. *	not Hot(air) *	

The second premise denies the consequent (*DC*) of the first. This inference is called *modus tollens*. Just like in an affirmation of the antecedent, the conclusion of the inference is true if the premises are true. We cannot say that for the first premise, so the conclusion may not be true.

To conclude the explanation his father further treats a possibly sufficient condition as a necessary condition. From the assumption that the sun rises when the air is hot he infers that when it is not hot, the sun also does not rise, thereby Denying the Antecedent (DA) of the first premise. This is also called an inverted *modus tollens*. This inference is invalid. Hence, the conclusion may be false even when the premises are all true.

P_9 If the air is hot the sun rises *	Hot(air) $\Rightarrow$ Rises(sun) *	
P_{11} In the evening the air cools down.	not Hot(air)	
		DA
P_{10} In the evening the sun sets *	not Rises(sun) *	

Assuming he translates not hot air (P_{11}) with cool air (P_{13}) and a not rising sun (P_{10}) with a setting sun (P_{14}), he rephrases this conclusion (invalidly) in the statement with P_{13} and P_{14} as premises: if the air cools down the sun sets (P_{12}), which given that the air cools would validly imply that the sun sets if the statement were true:

P_{12} If the air cools down the sun sets *	Cools(air) $\Rightarrow$ Sets(sun) *	
P_{13} In the evening the air cools down.	Cools(air)	
		AA
P_{14} In the evening the sun sets *	Sets(sun) *	

We can also interpret his whole explanation in yet another way. He could have assumed that the premise: if the sun rises the air is hot, stated that the rising sun is a necessary condition for hot air and hence infer that if the sun sets the air cools, via contraposition. But then he Affirms the Consequent (*AC*) of this proposition, also called an inverted *modus ponens*, to infer that if the air cools, the sun sets:

P_{15} If the sun sets the air cools down *	Sets(sun) $\Rightarrow$ Cools(air) *	
P_{13} In the evening the air cools down	Cools(air)	
		AC
P_{14} In the evening the sun sets *	Sets(sun) *	

By no great surprise this is invalid since the antecedent is not a necessary condition but a sufficient condition. In that case when the consequent of the first premise is known to be true the antecedent could be true, but could possibly be false just as well.

What can we conclude from this? Today, Calvin's father's explanation is gathered to be wrong. But is this because his hypothesis is unscientific, or because many of his inferences are fallacies? If we look at the beginning of modern science, three centuries ago, then what would we expect?

The Inquisition

In the seventeenth century Galileo Galilei defended the Copernican heliocentric theory. This theory put the sun at the center of the solar system, and explained that the sun sets because the earth turns on it own axis and revolves around the sun. It also explained the phases of Venus that Galileo first observed with his self made telescope.

Venus waxes and wanes as viewed from the earth, similar to the moon's phases. When Venus is full, we cannot see it because the sun is in the way. As Venus wanes from the full phase, it also gets bigger because it is approaching us. When it is closest to us, we cannot see it because no light is reflected towards us. This could be explained if it was assumed that both Venus and the Earth rotate around the sun. If you put the earth in the center then you could only explain it when you assumed Venus to rotate around the sun while Venus and the sun both rotate around the earth.

In 1616 Galileo was formally warned by the church to stop this defense. The reason for this censure was not that the claim was considered wrong, or that teaching so undermined the Church. Rather, it was claimed that Galileo's proof for the theory was not logically valid. Galileo's main argument depended on the fact that the theory explained why the planet Venus shows phases. Yet, he could not prove this deductively. The argument ran as follows:

> If the planetary system is heliocentric, then Venus will show phases.
> Venus shows phases.
> ___
> Hence, the planetary system is heliocentric

So the argument was based on an affirmation of the consequent, a fallacy well known by the Aristotelian clergy. While Venus does indeed show phases, the planetary system being heliocentric may not be the only condition under which that is true. The clergy pointed out the flaw and Galileo was ordered not to put forth this idea as proved.

Pope Urban VIII, who just as Galileo was a member of the Academy of Lynxes, a scientific society formed in 1603, informally lifted these orders in 1633. There is evidence that the Pope gave Galileo the opportunity to neutrally compare the heliocentric theory with the geocentric system of Ptolemy, and come up with a deductive proof.

But in the book he then wrote he patronized the Pope, who was greatly offended. As a result Galileo was accused of disobeying the order of 1616 to stop his defense of the Copernican system. Even though Galileo could produce a letter that showed he was merely warned instead of ordered, he was threatened by the Inquisition, shown the "instruments" (of torture), and sentenced to house arrest for the rest of his life. Hence, it was disobeying orders to stop using a fallacy that got him convicted by the Inquisition, and not committing heresy, since technically the Copernican system was never declared heretical (Gingerich, 1992). However, today science accepts Galileo's explanation. But is this because his reasoning is scientific?

4.3 Induction

A typical scientific explanation can never deductively follow from what we already know or have observed, because most scientific hypotheses include assumptions and predictions about future or other not observed situations. It is logically always possible that those situations will be different.

In his defense of the Copernican system Galileo not only needed to defend a scientific theory, but also a manner of reasoning. Galileo employed an inductive inference. The conclusion of an inductive inference can contain more or other information than its premises, hence it is not deductively valid. Deductive inference preserves the truth of its premises so as to encompass its conclusion, while an inductive inference expands beyond them.

However, scientific reasoning is not void of deductive reasoning. Logicians consider a sound explanation to be a deductive conclusion from a number of true hypotheses. The problem with scientific hypotheses is that you can never know for sure whether they are true. The philosopher Karl Popper stressed that what you validly can know about a hypothesis is that it is false. If a hypothesis claims that all particulars of a type have a property, then only one particular of the type without that property will validly imply that the hypothesis is incorrect.

At the beginning of this century, the philosopher Charles Sanders Peirce coined the term 'abductive inference' to distinguish Galileo's inference from other kinds of inductive inference like generalization (GEN). With generalization you infer that if a number of particulars of a type have a property, then all particulars of that type have that property. So for example:

The fact that a number of particulars of type A have property C is observed;

Hence, there is a reason to suspect that all A have property C

According to Peirce the function of abduction is *ampliative*, to introduce new ideas. A hypothesis suggested by abduction should contain predictions about other properties or other types of particulars as well. In his later work Peirce (1958, 5.188) put forward the following often quoted definition of abductive inference:

"The surprising fact, C, is observed;
But if A were true, C would be a matter of course.

Hence, there is a reason to suspect that A is true."

Abductive inference is actually part and parcel of everyday common sense reasoning. But it seems that it can lead to the wildest of explanations, as Calvin can attest. But even though his father commits enough deductive fallacies to experience more of the "instruments" then just their sight, had he lived three centuries ago, his explanation is not problematic just because of its inductive nature. Both Galileo and Calvin's father seem to follow the same inference. But then what makes Galileo's inference differ from that of Calvin's father's?

4.4 Abduction

How does Peirce's definition of abduction compare to other kinds of inductive inferences? Let us take a closer look at Peirce's inference scheme and our examples. We will address the similarities and differences. The examples are summarized in Table 4.1. The properties Hot and Rises are abbreviated to H and R respectively.

Inference	Premise 1 (Observed C)	Premise 2 (If A where true C follows, if)	Conclusion (A)
GEN	H(sun) $\Rightarrow$ R(air)	$\varnothing$	$\forall$x H(sun)$\Rightarrow$R(x) *
DVS	H(air) $\Rightarrow$ R(air)	sun part of air	H(air) $\Rightarrow$ R(sun) *
CAU	R(sun)	H(air)	H(air) $\Rightarrow$ R(sun) *
AA	H(air)	H(air) $\Rightarrow$ R(sun) *	R(sun) *
DC	Not R(sun)	H(air) $\Rightarrow$ R(sun) *	Not H(air) *
DA	Not H(air)	H(air) $\Rightarrow$ R(sun) *	Not R(sun) *
AC	Cools(air)	Sets(sun) $\Rightarrow$ Cools(air) *	Sets(sun) *
AC	Phases(Venus)	Center(sun)$\Rightarrow$ Phases(Venus)	Center(sun) *

Table 4.1: Summary of examples of the discussed inference types

For an inference to fit Peirce's definition of abduction, Premise 1 should follow as a matter of course if Premise 2 and the Conclusion are both assumed to be true. If we compare the example inferences with this definition we notice that generalization (GEN), division (DVS) and causality (CAU) fit the definition well. If the conclusion and Premise 2 are true, then premise 1 is also true. If a generalization is true, then the truth of a particular follows as a matter of course. In the division example Premise 1 follows based on premise 2 and the assumption in the conclusion that the property of a whole is also a property of its parts. If the causal implication in a conclusion and premise 2 would be true, premise 1 would follow by *modus ponens*. In sum, the inferences GEN, CMP and CAU can be seen as special kinds of abduction, according to Peirce's definition.

The next two types in Table 4.1 do not fit the definition. Not remarkably these are the deductively valid inferences affirmation of the antecedent (*AA*) and denial of the consequent (*DC*). These will of course not fit a definition of an abductive inference. In abduction the conclusion is an explanation, in deduction the premises are. However, in the example Calvin's father used these inference kinds incorrectly, because he wrongly assumed the premises were true, to conclude the truth of the conclusion. Denial of the antecedent (*DA*) fits the definition well. The implication in Premise 2: if H(air) then R(sun) is logically equivalent to: if not R(sun) then not H(air). Not surprisingly the affirmation of the consequence (*AC*) most resembles the definition of abduction. The observed fact C affirms the consequent of A $\Rightarrow$ C, where A is the conclusion. Both the explanation of the sun set and the phases of Venus follow that inference.

However, there is an important difference between the two. Premise 2, the implication if A then C, is true in the case of Galileo but uncertain in the case of Calvin's father. The implications are actually of a different nature. One is itself a hypothesis and the other a logical consequence. The former consists of a so called material implication and the latter of a logical or semantic implication. Let us take a closer look at the nature of implication and its role in Peirce's definition.

Implications

A material implication is a conditional statement that connects two independent statements. These statements may be either true or false depending on other conditions. The material implication asserts that when the antecedent is true, the consequent will also be true. Because of this property of the conditional statement it is argued that the material implication can represent a causal relation between two events described by the antecedent and the consequent. However, the truth of the conditional statement can already be settled by the status of only one of its constituents. The material implication is by definition already true when the antecedent is false or the consequent is true. Let us look at Table 4.2 to follow this.

Antecedent	Consequent	A $\rightarrow$ C
True	True	True
False	False	True
False	True	True
True	False	False

Table 4.2: Truth table of the material implication

In Table 4.2 I summarize all possible truth value combinations of a material implication. The material implication is false only if the antecedent is true and the consequent is false. The statement "if the air cools down the sun sets" is such a statement. It states that it will not happen that the sun does not set while the air does cool down (row 4). It still allows for the possibility that the sun sets even though the air does not cool down (row 3).

Now let us look at Galileo's statement. If the planetary system is heliocentric, then Venus will show phases. This implication differs in nature. The antecedent statement logically implies the consequent statement, and many others. For instance it will also imply under what conditions the sun will set. When you say that the antecedent is true, you say that all its consequences are true, by implication. Formally we say that all models that make the propositions true that make up the antecedent of a semantic implication, will also make the consequent true. The models of the antecedent constitute a subset of the models of the consequent. As a notation we will, following tradition, use $A \models C$ for semantic implication, and $A \rightarrow C$ will denote material implication. For the language of predicate logic it has been proved that if C is semantically implied by A, it can also be deduced from A, written as $A \vdash C$, and vice versa.

Another important difference between material and semantic implication is shown by the set of the inferences that each allows. Given that C is true, it can be inferred that $A \rightarrow C$ is true, but you cannot infer the truth of A. Yet given that C is true, it *cannot* be inferred that $A \models C$ is true, but you can say that A is confirmed. However, this is only the case if $A \models C$ is true. Even if you assume that $A \rightarrow C$ is true, but $A \models C$ is not, then C does not confirm A. If it is known that $A \models C$ is true, you can (non deductively) infer the truth of A if all its consequences are confirmed.

Let us return to the definition of abduction. Apparently Peirce meant "if A were true then C would be a matter of course" to be a semantic implication. Abduction based on a semantic implication will introduce a hypothesis that may have many other implications. Hence the use of the term abduction: it forces alien statements into the explanation.

A generalized material implication, such as $\forall xy(\mathrm{A}(x) \rightarrow \mathrm{C}(y))$, may also entail new predictions, but they are usually about the same properties, A and C, and the same kind of objects, all x and y such that $\mathrm{A}(x) \rightarrow \mathrm{C}(y)$. The antecedent of a semantic implication, such as $A \models C$ may entail predictions about different properties and objects as well.

Galileo's inference was based on a semantic implication and Calvin's father assumed a material implication. But are abductions based on material implications unscientific? If that where so then we could not use laws to explain phenomena. The material implication "if the atmosphere pressure drops the air will cool down" could then not be used to explain why the weather cools down. Even Galileo's explanation of the phases of Venus would run into trouble. His hypothesis entails many material implications as possible consequences, *e.g.* (using abbreviations):

C: {position i of the sun and Earth $\rightarrow$ phase j of Venus}
A: {Center(sun)} $\models$ C: {position i of the sun and Earth $\rightarrow$ phase j of Venus}

A: {Center(sun) *}

To explain a particular phase of Venus an abduction could infer a particular position of planets. That would not necessarily need a semantic implication to be scientifically acceptable. A law could be formulated that relates the position of the sun and Earth to the phases of Venus, that could explain a particular phase on the basis of a particular position:

$$\frac{\begin{array}{l}\{\text{C: phase i of Venus}\}\\ \{\text{A: position x of the sun and Earth} \rightarrow \text{C: phase y of Venus}\}\end{array}}{\{\text{A: position j of the sun and Earth *}\}}$$

In many scientific areas not much more is known than material laws. So it may be desirable for Peirce to infer a rich logical hypothesis, but a material implication is not unscientific by its nature.

Definitions

The main difference between affirming the consequence of a material implication and affirming the consequence of a semantic implication is a difference in category. The former is part of the latter. Let us call the former kind *material abduction* and the latter kind *semantic abduction*:

$$\frac{\begin{array}{l}C\\ A \rightarrow C\end{array}}{A}\ \text{material abduction} \qquad \frac{\begin{array}{l}C\\ A \models C\end{array}}{A}\ \text{semantic abduction}$$

To avoid confusion between the two I will adopt the following notation. I will use the propositions C, A → C, and A, etc. to talk about statements that a semantic abduction reasons about. The premises and conclusion of a semantic abduction are sets that contain these statements. The first will be a set called P, containing a proposition about the world; the second premise a set H containing the hypothesis statement(s) that together with background assumptions B implies P. I can now define the different kinds of abduction as follows:

Definition 1 *Semantic abduction.* A semantic abduction is an inference that affirms the consequent of a semantic implication (ACS). Given the antecedent B ∪ H that semantically implies P, the affirmation of the consequent P infers hypothesis H:

$$\frac{\begin{array}{l}\text{Proposition P}\\ \text{Background B} \cup \text{Hypothesis H} \models \text{Proposition P}\end{array}}{\text{Hypothesis H: }\{*\}}\ \text{ACS}$$

In this scheme the set containing only a star {*} denotes a set of propositions with unknown truth value. A semantic abduction can encompass different kinds of inductive inferences. Affirming the consequent of a material implication is just one special case.

Definition 2 *Material abduction.* A material abduction is an inference that affirms the consequent of a material implication (AC), as a special case of a semantic abduction.

$$\frac{\begin{array}{l}\text{Proposition P: }\{C\}\\ \text{Background B: }\{A \rightarrow C\} \cup \text{Hypothesis H: }\{A\} \models \text{P: }\{C\}\end{array}}{\text{Hypothesis H: }\{A^*\}}\ \text{ACS: }\{AC\}$$

The material implication A → C could either be part of the hypothesis or belong to the established background assumptions B which should then be part of the antecedent of the semantic implication. Affirming the consequent of a material implication (AC) is the typical example of a semantic abduction. But the other discussed inductive inferences can be an instance as well, *i.e.*: denial of the antecedent (DA); division, attributing properties of wholes to parts (CMP); inferring causality between co-occurring events (CAU); and generalization from particulars to groups (GEN); see Table 4.3.

Explanation (ACS)	Proposition P	Background B	Hypothesis H
AC	$C(y)$	$A(x) \rightarrow C(y)$	$A(x)$ *
DA	Not $A(x)$	$A(x) \rightarrow C(y)$	Not $C(y)$ *
DVS	$A(p) \rightarrow C(p)$	p part of w	$A(p) \rightarrow C(w)$ *
CAU	$C(y)$	$A(x)$	$A(x) \rightarrow C(y)$ *
GEN	$A(i_2) \rightarrow C(i_2)$	$A(i_1) \rightarrow C(i_1)$	$\forall x\ A(x) \rightarrow C(x)$ *

Table 4.3: Some examples of explanation as semantic abduction (ACS): given B ∪ H ⊨ P, proposition P affirms the consequent to infer H.

But if inferences with material and semantic implications are part and parcel of abductive reasoning then we do not have an reason why the hypotheses of Calvin's father and Galileo differ. When is an abductive inference a scientific explanation?

4.5 Formation

There are in fact two very distinct ways to understand the terms "abductive inference" and "scientific explanation". In the first way the term is a verb and in the second way it is a noun. In the former sense it refers to the process of inferring and explaining. In the latter sense it refers to the product of that process. Abductive inference as defined by Peirce is first of all a process of inference. You assume two premises, and the conclusion of the inference is an explanation that could be correct.

But how do you know what specific hypothesis to infer? You could logically infer many different possible hypotheses that all would imply a surprising observation. (Why does the furnace not work? Is the switch broken? Is the gas pipe fractured? Oh wait a minute, did I pay my bill?) And on the other hand, coming up with only a single explanation that would non-trivially imply all our observations is no trivial exercise. Peirce's abductive inference scheme tells us nothing about what specific hypothesis to infer. He said: "The abductive suggestion comes to us as a flash" (1958, 5.181). His scheme only tells us under what condition to infer a statement as a hypothesis.

In the 1930's the philosopher Hans Reichenbach (1938) suggested that logicians should only address the problem of the nature of scientific theories and of their evaluation. The search and formation of new theories was taken to be an erratic and non-rational process that was not open nor relevant for a logical inquiry of knowledge. He suggested a distinction between a context of discovery and a context of justification in the study of scientific knowledge. This served as a demarcation of the problems relevant for epistemology. The study of the formation and discovery of hypotheses should be a problem for psychology. So according to Reichenbach's claim, logic should be able to evaluate a scientific explanation, regardless of how a hypothesis was inductively inferred or conceived. A good scientific explanation should satisfy certain logical conditions. One of those we already encountered: an explanation should logically imply the surprising observation. By its definition we already are sure that an abductive conclusion satisfies that condition. But both the explanation given by Galileo and that given by Calvin's father do so. So the question remains: what other conditions make an explanation scientific?

4.6 Explanation

Philosophers of science have long thought about the nature of a good scientific hypothesis. They set up certain conditions that would mark a valid and potentially successful explanation. We saw that any proper explanation should deduce a proposition from the explaining assumption. This is, by definition, possible if H combined with background assumptions B semantically implies P.

The set P may contain particular propositions, such as the proposition that certain objects have certain properties at a certain time. It can also contain general propositions, such as the proposition that all objects of a certain kind have a certain property, or that some object will have a certain property at a certain time. The background set B and hypothesis set H may also contain both particular and general propositions. General propositions in H and B can imply another general proposition in P. Together with an assumption about a particular they can imply particular propositions in P. In empirical sciences explanations are sought for particular or general facts about the world that are observed or assumed to be true. We will use the set O to refer to propositions about the world that are regarded to be certain because they are observed, given some criterion of proper observation.

In philosophy of science several conditions for a proper scientific explanation are proposed (*cf.* Aliseda-LLera, 1997, Flach, 1995). We will introduce some of them. There are both conditions for the explaining hypothesis and for the explained proposition. Given background assumptions B, proposition P, observations O; hypothesis H properly explains P if:

Conditions for the explaining hypothesis H:

HC_1. Implication:	$B \cup H \models P$
HC_2. Consistency:	H is compatible with B
HC_3. Non-triviality:	$H \not\models P$
HC_4. Simplicity:	H is minimal among the H's were $B \cup H' \models P$

Conditions for a proposition P that needs to be explained:

PC_1. Observation:	P is assumed to be true
PC_2. Novelty:	$B \not\models P$
PC_3. Anomaly:	$B \models \text{not } P$
PC_4. Indifference:	$B \not\models P$ and $B \not\models \text{not } P$

If PC_1 and any of the conditions PC_2 to PC_4 hold for a proposition, a hypothesis is required for which all conditions HC_1 to HC_4 hold. These are considered to be ideal conditions, proposed and defended by different logicians. Let us go through them and at the same time see whether the heliocentric hypothesis of Galileo and the hot air hypothesis of Calvin's father satisfy them:

Heliocentric hypothesis:	H: $\{\text{center(sun)}\} \models \{\text{phases(Venus)}\}$
Hot air hypothesis:	H: $\{\text{air cools} \rightarrow \text{sun sets}\}$

We already encountered the first condition HC_1. It dictates that an explanation of P consists in a deductive inference of P from B and hypothesis H. The philosopher Carl Hempel (1965) calls this hypothetical-deductive inference. By this condition an explanation consists of either a denial of the consequent of a hypothesis (DCH) or an affirmation of the antecedent of a hypothesis (AAH):

$$\frac{\begin{array}{l}\text{Background B: }\{A\}\\ \text{Hypothesis H: }\{A \rightarrow C\}\end{array}}{\text{Proposition P: }\{C\}}\ \text{AAH} \qquad \frac{\begin{array}{l}\text{Background B: }\{\text{not } C\}\\ \text{Hypothesis H: }\{A \rightarrow C\}\end{array}}{\text{Proposition P: }\{\text{not } A\}}\ \text{DCH}$$

In this way if B and H are true then they explain P. If P is true then it confirms H assuming B. Both the heliocentric and the hot air hypotheses comply as we saw earlier in our discussion in Section 4.4.

The second condition (HC_2) dictates that implications of $B \cup H$ should not contradict each other. That means that in case of contradiction either H or B should be substituted by a different set of propositions. The implication of the hot air hypothesis appears consistent with our other assumptions. However, the Heliocentric hypothesis contradicts the assumptions of Ptolemy, which were part of the background knowledge that, in Galileo's time, was assumed to be true. Condition HC_3 is meant to prevent the use of ad hoc hypotheses. It dictates that an observed proposition should not solely follow from the hypothesis. It should at least depend on some other assumptions that are not purely hypothetical. Both hypotheses comply. The fourth condition makes some requirements about the complexity of the hypothesis, given some interpretation of “minimal”. Both hypotheses do not seem unnecessarily complex.

In the next part of this thesis when we look at scientific practice, we will see that usually no employed hypothesis complies with all four conditions. It is usually argued that this fact does not mean that those hypotheses are unscientific or that the conditions are wrong. It is rather argued that the conditions define an ideal to be approached by science, given some justification for the conditions.

Now let us turn to the conditions for the explained proposition. Condition PC_1 states the assumption that a hypothesis in empirical science explains observations. If a consequence of a hypothesis is not observed, or on some other grounds certain to be true, then there is nothing to explain. While the four conditions for a hypothesis are each of them desirable, conditions PC_2, PC_3 and PC_4 are disjunctive; only one needs to apply. PC_2 states that a proposition only needs an explanation by a hypothesis H if it is not implied by what we already assume. PC_3 states that the observed proposition is in contradiction with the implications of our earlier assumptions. Or the background could be totally indifferent about it, as stated by PC_4.

The phases of Venus were a real anomaly (PC_3) for the assumption of Ptolemy. So by these conditions it required an explanation, which was properly provided by the heliocentric hypothesis. Yet, together with the assumption that the earth evolves around its axis, the rising of the sun is already explained by that hypothesis. It did not need another explanation. But logically there are always more explanations possible. So again, what makes the former a better explanation than the latter?

4.7 Prediction

Karl Popper contended that an explanation is no scientific explanation if it cannot be tested. He maintained that, before anything else, scientific reasoning is the systematic search for errors in our assumptions. Peirce also argued that therefore a proper explanation should at least predict propositions that are either novel, anomalous, or indifferent with respect to current (theoretical) assumptions. It should predict a P that satisfies conditions PC_2, PC_3, or PC_4, but not PC_1. Many wrong hypotheses may explain given observations, but true hypotheses will always correctly predict a new unobserved fact.

Logically a prediction of a proposition can be considered to be the same as an explanation, it should deductively follow from the hypothesis and background assumptions. But just as in the case of the definition of abduction we can make a distinction between affirming the antecedent of a material implication (AAH) or of a semantic implication (AAS). The former can again be part of the latter:

Definition 3 *Semantic prediction.* A semantic prediction is an inference that affirms the antecedent of a semantic implication (AAS). Given the antecedent $B \cup H$ that semantically implies P, the affirmation of the antecedent infers prediction P. Affirming the antecedent of an hypothetical material implication (AAH) is the prototypical example:

Background B: {A}
Hypothesis H: {A $\rightarrow$ C}
B: {A} $\cup$ H: {A $\rightarrow$ C} $\models$ P: {C}
———————————————————— AAS: {AAH}
Proposition P: {C}

We can consider the affirmation of the antecedent of a semantic implication as the general definition of prediction. Affirming the antecedent of a hypothetical material implication (AAH) is the prototypical AAS that provides the best bait for catching the truth value of an hypotheses by testing its prediction in the pond of nature. It is the ace of hypothesis testing. But others can be possible as well. A complete typology would be:

AAH: affirming the antecedent of a hypothesis
DCH: denying the consequent of a hypothesis
DAH: denying the antecedent of a hypothesis
ACH: affirming the consequent of a hypothesis

HAA: hypothetically affirming the antecedent of a background assumption
HDC: hypothetically denying the consequent of a background assumption
HDA: hypothetically denying the antecedent of a background assumption
HAC: hypothetically affirming the consequent of a background assumption

The value of a prediction for a hypothesis can be measured by the information we gain if we find out that the prediction comes true. We can call this its strength. In case of AAH a background assumption affirms the antecedent of a hypothetical implication. One infers the strongest prediction, its truth value either confirms or refutes a hypothesis. It is also possible to hypothetically affirm the antecedent of a hypothesis in the background assumptions (HAA). This is weaker because if the prediction P is observed it will not inform you about the truth of the hypothesis. But if not P is true it will refute the hypothesis, see Table 4.4 for all types.

Prediction (AAS)	Background B	Hypothesis H	Prediction P	If P is true then H is?	If P is false then H is?
AAH	A	A $\rightarrow$ C	C	Confirmed	Refuted
ACH	C	A $\rightarrow$ C	A *	Confirmed	Confirmed
DCH	Not C	A $\rightarrow$ C	Not A	Confirmed !	Refuted
DAH	Not A	A $\rightarrow$ C	Not C *	Confirmed !	Confirmed
HAA	A $\rightarrow$ C	A	C	?	Refuted
HAC	A $\rightarrow$ C	C	A *	Confirmed	?
HDC	A $\rightarrow$ C	Not C	Not A	?	Refuted
HDA	A $\rightarrow$ C	Not A	Not C *	Confirmed	?

Table 4.4: Types of prediction (AAS) of different strength: Given B $\cup$ H $\models$ P, background B affirms the antecedent of hypothesis H to infer prediction P.

So the route from theory to experiment is determined logically by an informative prediction that can be tested. The strongest test, the one that provides the most information, is always preferable. But there can be pragmatic problems to test it. The first problem is whether it is possible to observe the predicted property of a phenomenon. If not, the prediction is useless as an empirical test for the hypothesis. Most effort in the defense of the heliocentric hypothesis for Galileo was put in constructing a strong enough telescope to observe the predicted phases of Venus.

Some predictions state a possibility that will not naturally occur. But can you create an intervention such that the initial conditions for the possibility are forced? This is not always possible. The latest technology often makes observations and interventions possible that lay beyond our reach or sight without it. This makes technology an epistemological factor. Other predictions can easily be observed but will never occur according to the hypothesis. If they do not, how will you know they never will? Here lies the main problem of the hot air hypothesis.

P_9 If the air is hot the sun rises *	Hot(air) $\Rightarrow$ Rises(sun) *

This hypothesis logically implies that either:

P_{16} The air cools and the sun sets	Cools(air) & Sets(sun)
P_{17} The air is hot and the sun rises	Hot(air) & Rises(sun)
P_{18} The air cools and the sun rises	Cools(air) & Rises(sun)

This is consistent with all our observations. But it also implies that it will never be so that the antecedent is true and the consequent is false, *i.e.*:

P_{19} The air is hot and the sun sets	Hot(air) & Sets(sun)

This is its only test opportunity, that is unobserved so far. So, the only way to test the hypothesis is to create a situation where the air is kept hot by an intervention, and wait for the sun not to set. But how can we do that? The hypothesis is testable in theory, but not in practice. But does that make it an unscientific hypothesis?

4.8 Comparison

According to Theo Kuipers (Kuipers 2000) the question about the rationality of scientific reasoning is not only what it means to have a good scientific explanation, but also what it takes to have a better one. In this approach it is evaluated how one explanation compares to another. The best hypothesis would imply all true propositions about a domain. But acknowledging that this is the ideal goal, the value of an hypothesis is measured by how far it might be away from that goal in comparison with another hypothesis. A hypothesis that includes more true propositions then a competitor and has less counterexamples might be closer to the truth. This intuition is formalized in a rule of success. This inference rule is not deductive in nature, but abductive. If the more successful theory would be closer to the truth that would explain why it is more successful. In this light Calvin's father's explanation is not so much unscientific, but just not as good as Galileo's, because next to explaining the phases of Venus, it also explains other phenomena such as stellar parallax. Yet there are more conditions formulated that characterize a good scientific explanation. In Chapter 6 I discuss how one of them, the simplicity of a theory, is related to the probability of its predictions.

4.9 Conclusion

In this chapter I asked the general question: what is the rational use of theory and experiment in the process of scientific discovery, as proposed in the study of logic? More specifically I looked at logical prescriptions for scientific theories and scientific reasoning. To address these topics I discussed an illustrated example that contains a series of inferences that are marked as fallacies from the viewpoint of logic and argumentation. Yet I argued that these inferences are common in science and part of abductive inference as defined by C.S. Peirce. I further made a category distinction between semantic abduction and material abduction. I argued that the latter, as well as other types of inductive inference, constitute a special type of the former under this definition.

I first discussed the validity of deduction, induction, and more specifically abduction in scientific reasoning. Scientific reasoning includes inferences about hypotheses of which we do not or cannot know whether they are true. What logic tells us most importantly is what a valid inference looks like. It defines under what conditions we can safely accept the conclusion of an argument. In the case of deduction we know that the conclusion is true when the premises are true. In the case of abduction or explanation we can know that the premises are true, but we have no guarantee for the conclusion. What valid reasoning can do is check whether the conclusion of an inference satisfies certain conditions. For explanation it can check whether a hypothesis is *e.g.* successful, non-trivial or consistent. But these are ideal conditions that still do not determine its truth. Yet they may be functional for establishing its similarity to the truth. I argued that prediction is not just deduction. A good prediction with the aim to test a hypothesis should satisfy other conditions as well.

In sum, what is rationality in scientific discovery? According to logic scientific discovery is a process of observing, describing, explaining, predicting and intervening in natural phenomena. A phenomenon is empirically discovered by observing it in the world. An explanation of that phenomenon may predict the existence of other phenomena that could be observed or created by a specific intervention in an experiment to test that prediction. As an answer the specific questions of this thesis from Section 1.3, we may not that according to studies in logic the following holds:

Question 1 What is the structure of a scientific theory? Theories are logically represented as a set of hypothetical propositions H that together with propositions describing background assumptions B semantically imply the propositional facts P they explain, *i.e.* $B \cup H \models P$.

Question 2 What is the process of scientific reasoning? The process of reasoning is different for the explanation and prediction of facts, see Table 4.5.

Problem	Premise	Background	Inference	Conclusion	Properties
Explanation	P	B	Abduction	H: {*}	$B \cup H: \{*\} \models P$ H is minimal
Prediction	H	B	Deduction	P: {*}	$B \cup H \models P: \{*\}$ P is informative

Table 4.5: Short overview of the inference types discussed in this chapter

Explanation of a phenomenon involves the abduction of a simple hypothesis from which the properties of an observed instance of that phenomenon can de deduced. Induction, as conceived as the generalization from the property of one instance of a category to all instances, is in this sense a special kind of abduction. Prediction involves the deduction of informative consequences from a given hypothesis.

Question 3 What is the route between theory and experiment? The route between theory and experiment typically involves six steps (explanation follows):

1. Observation of a phenomenon P: observe p_m and p_n
2. Description of P: $P: \{A(p_m) \rightarrow C(p_n)\}$
3. Explanation of p by a new hypothesis: $B \cup H: \{*\} \models P$
4. Prediction by a hypothesis: $B \cup H \models P: \{A(p_i) \rightarrow C(p_j) *\}$
5. Intervention in an experiment: create $A(p_i)$
6. Observation in an experiment: observe p_j

An observation of a phenomenon p in step 1. consists in observing natural objects such as *e.g.* p_m and p_n. The description of p in step 2. consist in categorizing the properties of the phenomenon, *e.g.* in A and C, and making a statement about those properties, *e.g.* $A \rightarrow C$. After finding an explanation, in step 3., that implies that statement, a prediction could be deduced in step 4. This prediction can include that if an object p_i has property A, then object p_j will have property C. In step 5. the situation $A(p_i)$ can be forced by an intervening experiment. The last step, observing the consequence of the intervention, closes the circle by being of the same kind as the first step. The experimental discovery of the truth value of the prediction either refutes or confirms the hypothesis (or a background assumption). A more advanced logical approach can evaluate an hypothesis by comparing its success with that of competing hypotheses.

In the next chapter I will discuss rationality in the process of scientific discovery in terms of the study of cognition. In this approach rationality can be understood as part of learning to solve problems heuristically.

* * * *

Chapter 5

Cognition

5.1 Introduction

In cognitive science, rationality in scientific discovery itself is being studied as an interesting cognitive phenomenon. One popular view is taking scientific discovery as just a form of human problem solving (Langley et al. 1987). One of the most successful theories about human problem solving is developed by John R. Anderson (Anderson 1993, Anderson & Lebiere 1998). It is called ACT-R, meaning Adaptive Control of Thought – Rational. The ACT-R theory deals with the cognitive mechanisms of learning and rational behavior. It aims to explain how people make an assumption or take an action to observe or change something in the world, in such a way that the probability to achieve a specific goal is high and the cost of time to achieve it is low. ACT-R is implemented in a computer program to test the performance of specific models of problem solving strategies.

The general question of this chapter is: what is rationality in scientific discovery, according to the psychological study of cognition? As a general model of human cognitive abilities, ACT-R should also be able to model specific cognitive processes involved in scientific problem solving. In this chapter I investigate how it could do that. The particular question that is answered in this chapter is: how can one understand and model scientific discovery with ACT-R?

I will first, in section 5.2, introduce a distinction between primary and secondary epistemology. Analogously to these types I make a distinction between primary (or native) and secondary (or acquired) processes of cognition. I will use this distinction to discuss how beliefs, goals and search methods are created, selected and evaluated according to the ACT-R theory in section 5.3 to 5.5. In section 5.6, I discuss how scientific discovery, as modeled in Simon and Langley's BACON.1 (Langley et al 1987) and Thagard's PI (Thagard 1988) can both be modeled in ACT-R as similar forms of abductive inference. I demonstrate and discuss how ACT-R's primary mechanisms nicely subsume PI's hypothesis evaluation process. Then, I discuss BACON.1's search methods and how they can be learned by analogy from examples. In section 5.7 I discuss the nature of theory and method in the different models. 5.8 discusses the difference between the logical and psychological views on explanation and prediction. I end this chapter in section 5.9 with a discussion and general conclusion, answering the specific questions from section 1.3.

5.2 Primary and secondary

The claim that philosophy of science can learn something from cognitive psychology is endorsed by the philosopher Alvin Goldman. He argues that epistemology, the study of justified belief, should take explicit account of empirical studies of cognitive processes (Goldman 1986). Among the many factors that influence the forming of belief he distinguishes basic cognitive processes from acquired belief forming methods.

The first category, *basic processes*, include processes of perception, memory, attention, concept formation, problem solving, learning and reasoning. Goldman argues that these natural or native processes are suitable objects for normative epistemic evaluation, and comprise the domain of *primary* epistemology. *Secondary* epistemology comprises the normative evaluation of acquired belief forming *methods* like algorithms, techniques or procedures. A method can either be a general, topic neutral, or a task specific procedure for arriving at beliefs.

In forming a belief, basic processes and methods are intrinsically intertwined. When someone needs to solve a problem and several methods are available, the basic processes determine which method is applied, and also which new methods are created or added. So evaluating a resulting new belief depends on the reliability of both the basic processes and the specific applied method.

So in short, primary epistemology is concerned with the evaluation of basic, *i.e.* native or natural, cognitive processes, and secondary epistemology is concerned with the correctness of acquired belief forming methods. To explain how such processes and methods are explicated in the ACT-R theory, I will first use Goldman's distinction to differentiate between two general types of cognition, *i.e.* primary and secondary cognition.

By *primary cognition* I mean native or basic cognitive processes and structures, whereas by *secondary cognition* I mean acquired cognitive processes and structures. In this way we can also distinguish acquired *structures*, like beliefs and goals, from basic structures, like the memory activation values used by basic or primary cognitive processes in ACT-R.

5.3 Declarations and procedures

Anderson's ACT-R explains human (problem solving) behavior as the result of acting according to two types of knowledge: declarative and procedural knowledge (Anderson 1993). Declarative knowledge consists of declarations of beliefs and goals, and resides in a person's declarative memory. Procedural knowledge consists of procedures that can create and modify a persons beliefs and goals. It contains our cognitive skills, or our *know how*. In ACT-R declarative knowledge is represented as a collection of memory structures called *chunks.* A chunk is an abstract representation of a belief or goal structure. Its basic elements consists of a list with slots and slot values. For example:

```
(Johannes_Kepler
      ISA                 person
      BORN                "27 December 1571"
      PROFESSION          scholar
      ACHIEVED            "discovery laws of planetary motion"
      FEARED-MOST         "invasion by the Turks"
      ETC                 ...)
```

The ISA ('is a') slot value represents the type of the chunk, and can be seen as a concept type name. Every concept type has the same slot-names, or concept attributes. So in our example, a person is something with a date of birth, a profession, etc. A slot value can in its turn also be a chunk. In this way declarative knowledge is structured in a network of memory chunks. In our example:

```
   (scholar
      ISA                 profession
      ACTIVITY            research
      ETC                 ...)
```

Procedural knowledge, or know how, is represented by production rules, or productions for short. Such a rule consists of a set of conditions and actions. The conditions, or left hand side (LHS), of a production can match with memory chunks which satisfy given constraints. When a matching succeeds, certain actions can be performed which are specified in the action, or right hand side (RHS), of a production. For example:

```
   (SUBTRACT
      =goal>
         ISA subtract
         VAR1             =x
         VAR2             =y
         ANSWER           nil
      =addition-fact>
         ISA addition-fact
         ADDEND1          =y
         ADDEND2          =z
         SUM =x
   ==>
      =goal>
         ANSWER           =z)
```

This production uses declarative knowledge of an addition fact to find the answer for a subtraction problem. A string with an '=' sign is a variable that is bound to a value by matching a chunk. The LHS, before the arrow, matches against any goal of which no answer is known and a fact (an addition fact in the example) that satisfies the values =x and =y of the goal slots. In the RHS, after the arrow, the found value =z of the addition fact is added to the ANSWER slot of the subtract goal.

In summary, this is what ACT-R poses that human problem solving is all about: matching productions (skills) to memory chunks (beliefs and desires). We can say that the chunks and productions themselves, constitute secondary cognition. A memory chunk is an acquired structure, a production is an acquired process. However, the processes that ACT-R really is about are the (native) mechanisms about *how* and *what* memory chunks and productions are used in problem solving.

In human problem solving often several (possibly mutually inconsistent) belief chunks can match a production's LHS. And for a given problem goal more than one production may apply. The ways the cognitive mechanism efficiently evaluate alternative chunks and productions constitute the main aspects of primary cognition.

5.4 Structures and processes

In Table 5.1, I summarize the main cognitive mechanisms according to the ACT-R theory, explicating their primary processes and structures. In the process of problem solving, (secondary) knowledge, containing of chunks and productions, is created, selected and evaluated by (primary) learning mechanisms. (This section discusses the ACT-R architecture up to version 3, primarily based on Anderson (1993).)

Cognitive mechanisms	Primary processes	Primary structures
Creation of chunks by:		
Concept-formation	(Specifying chunk types)	(Basic types?)
Perception	Specifying (new) chunks	(Constraints?)
Productions (RHS)	Specifying RHS chunks	-
Selection of chunks by:		
Productions (LHS)	Matching LHS chunks	-
Goal focus	Goal stack control	-
Activation	Preferring high $A_i=B_i+S_jW_jS_{ji}$	Value A_i
Base-level activation	Computing & learning B_i	Value B_i
Salience strength of j to I	Computing & learning S_{ji}	Value S_{ji}
Association of i with j	Computing W_j	Value W_j
Evaluation of chunks by:		
Activation	Preferring highest A_i	Value A_i
Creation of productions by:		
Analogy	Generalizing example chunks	Special slots
Selection of productions by:		
Goal focus	Matching LHS to goal focus	-
Chunks	Matching LHS to chunks	-
Matching time (latency)	Preferring low $T_p = S_i e^{-(A_i + S_p)}$	Value T_p
LHS chunks activation	Computing & learning A_i	Value A_i
Strength of production	Computing & learning S_p	Value S_p
Eval. of productions by:		
Expected gain	Preferring high value $E = PG - C$	Value E
Probability of success	Computing $P = qr$	Value P
Prob. of intended effect	Computing & learning q	Value q
Prob. of suc. after firing	Computing & learning r	Value r
Value of the goal	Specifying value G	Value G
Cost of production	Computing $C = a + b$	Value C
Cost of firing production	Computing & learning a	Value a
Cost of actions after firing	Computing & learning b	Value b

Table 5.1: Primary aspects of ACT-R's cognitive mechanisms

Table 5.1 summarizes the primary aspects of ACT-R's cognitive mechanisms (version 2.0). In the first column I list different kinds of primary cognitive mechanisms. These essentially control the creation, modification, selection and evaluation of secondary cognitive structures (memory chunks) and processes (productions). The primary cognitive mechanisms consist of primary processes (column 2), guided by, and modifying primary structures (column 3). I will discuss them briefly in the following subsections.

Creation

In the ACT-R theory, memory is ordered by *types* of memory chunks. A concept like 'person' in the example above, is supposed to have a given template of attributes. Every instantiation of a concept shares the same attribute slot names, but differs in their values. If you want to add something to memory, a concept type is necessary. But how do concepts come about in ACT-R?

In any cognitive creation or modification process we can make a distinction between the process that actually makes the creation or modification and that *what* is created or modified. In connectionist theories of cognition we often see that both are the same, that the concept creation process 'decides' on the concept types 'on the run'. In ACT-R there is no primary process specified that creates types, and the theory is silent about what types there should be. The modeler has to define them up front. Chunk types can also not be created or changed by learned productions, while chunk type instantiations can. So it is not clear whether we can consider concept types as primary or secondary structures, and if there are any basic constraints, or even basic or native types (like Jerry Fodor suggests).

The process of *perception* can add new chunks to memory. Again we can say that in ACT-R the process of adding them is a primary process. Yet how perception is constrained by concept types, or guided by problem solving is not defined in ACT-R, but in the perceptual/motor extension of the theory ACT-R/PM. I will not go into this extension here (see Anderson & Lebiere 1998).

Finally *productions* can add and modify memory chunks. That is what ACT-R is (mostly) all about, how and which productions modify and add chunks to memory. Once a chunk is added it will never be deleted. Its worst fate is never to be recalled. How, and what chunks are recalled is governed by processes of chunk selection. Productions themselves, as representations of learned skills, can only be created and added by a primary process of analogy. To connect actions to conditions, ACT-R starts out with a declaration of a problem example and its solution. When another problem of the same type is encountered, analogy will generalize a solution strategy from the known example. How that process works is discussed in the next section.

Selection

In a process of problem solving the selection of relevant chunks and productions is constrained in several ways. The main guiding mechanism of problem solving in ACT-R is *goal focus*. Goal focus is a kind of pointer to a chunk saying, "this chunk represents the goal I want to achieve", which in ACT-R means "that is the chunk a production should match with". ACT-R does not say how goal focus is initially specified. How a person is motivated to desire the accomplishment of a goal, however, is determined rationally in ACT-R. After setting the first goal, several primary and secondary processes influence how to achieve that particular goal by specifying and focusing on subgoals. The action, or right hand side (RHS) of a production can shift focus to another goal, which is implemented by a *push* of a new goal on *stack*. When a production has achieved the new goal, it can *pop* it from the stack, thereby changing focus to the next goal below it on the stack.

When an initial goal is set, ACT-R first selects a set of potential productions that can match with it. For a production to match, the given goal must be the first part of the production's condition or left hand side (LHS). An LHS usually contains other chunks which should match as well, given specified constraints, and need to be retrieved from memory.

ACT-R also models *latency*, which is the time it takes to match a production to memory and perform the action. How long that takes depends on the activation of the chunks needed. The latencies in the model should reflect the latencies in reaction time of subjects, measured in psychological experiments.

Activation is a basic property of every chunk. A chunk's activation value is the result of its prior base level activation plus the contribution of chunks that are part of the current goal context. This value increases with use. A primary learning process increases the association between two chunks every time they are both needed to solve a problem. According to Anderson, a chunk's activation denotes its posterior (logarithmic) odds that it will be needed in a given context, and the learning process is supposed to give the best estimate of that chance. When a chunk is not used its activation decays logarithmically. When it drops below a certain threshold, it can no longer be retrieved in the current context. Another context might however contain the right cues to boost the activation above the threshold again, re-enabling retrieval. Next to chunk activation, a production's strength also controls production selection. A production's strength increases after use, and is learned accordingly. Again its strength denotes its (logarithmic) odds of being needed.

So in sum, when focus is set to a goal, primary processes in ACT-R start to select productions that can match with it. A set of alternatives is gradually selected, depending on the activation of chunks in the productions' LHS, and the strength of the productions.

Evaluation

When several chunks can match a production's LHS, the chunks with the highest activation will be used. However, that is not the case for productions. Next to the time it takes to retrieve relevant productions, other primary evaluation processes contribute to determine what production will determine the next action.

During selection, potential productions are evaluated simultaneously by a primary process of rational analysis. This process diagnoses whether a given production is worth it to be fired. In order to do so it takes three estimations into account: the probability the production will be successful ($P = qr$); the value of the goal that is desired (G); and the cost of firing that production ($C = a + b$). A production's probability of success is a product of the probability of its intended effect (q) and the probability of achieving the goal after having achieved intended effort (r). The cost of a production is the result of adding the cost of the cognitive effort to fire the production (a) with the cost of actions needed to reach the goal after firing the production (b).

For example, if your goal is to lessen your thirst, and you are in front of a coffee machine, a production may be evaluated that urges to throw a coin in the machine to get a cup of coffee. Now q is the estimation that the machine will indeed return a cup, and r is the chance that only one cup will quench your thirst. The quantity a denotes the effort of putting in a coin, while b stands for the effort of emptying the cup. The quantities q and a can be estimated by repeated applications of the production. For example, if the machine is old and failed a number of times in the past, q will be low.

The quantities b and r are more difficult to estimate because they may refer to yet unknown actions. Anderson's solution is to base their estimates on how much the state achieved by the production differs from the desired goal. If the action of putting in a coin fails to provide you with a coffee, it is less likely that you will quench your thirst (r') and more effort will be needed to get a drink (b'). And in general the more effort already spent, the less likely you will achieve your goal at all, so the lower the probability (r').

The production with the highest estimated gain E (= PG-C) of the selected productions is generally preferred. In this way when the value of a goal or the probability of its success is high, the cost of a production plays a less important role. When you know the coffee machine often fails and is situated on another floor of the building, the cost of walking to it may not be worth one's while. But when you are really thirsty the cost loses out to the value of the goal. The best production rule given its PG-C is not always selected, but it has the highest chance of being fired.

When a production finally fires, its RHS or action side will be executed, changing beliefs or goals, or initiating hand an eye movement, like looking for the slit on the coffee machine and putting a coin in it. After firing, a new (sub)goal may be set by the production or from the goal stack, and the process of selecting, evaluating and firing a production starts all over. ACT-R stops when the initial goal is achieved and popped from the goal stack.

5.5 BACON and PI

In this section I discuss two computational models of scientific discovery, and how the structures and processes of these models can be modeled in ACT-R. Typical scientific problems are searching and evaluating descriptions and explanations for interesting observations. Herbert Simon and Paul Thagard proposed different explanations about how scientists (could) solve those tasks. They both modeled their theory in computer programs, respectively called BACON and PI.

The first of the BACON programs models the search for simple quantitative laws that describe the numerical data of observations, like Kepler's third law of planetary motion and Boyle's gas law. PI searches and evaluates qualitative explanations, like the explanation of the propagation of sound from its being a wave.

In PI, new hypotheses are searched and evaluated through a primary process of abduction and inference to the best explanation (IBE). In this section I will argue that abduction is better thought of as a secondary *acquired* process in ACT-R, generalized from examples by analogy, while IBE is subsumed by ACT-R's primary processes. I will further demonstrate that the heuristic search method for laws as implemented in BACON.1 can also be learned from examples.

Simple abduction in PI

Paul Thagard's theory of cognitive inductive processes, modeled in PI (processes of induction), includes several forms of abduction. I will consider its simplest form. Abduction, as discussed in Chapter 4, is a form of inductive inference. It is inductive in the sense that the truth of the conclusion of the abductive inference does not follow from the truth of the premises. As stated in Chapter 4, Peirce defined abduction as follows:

(P_1) "The surprising fact, C, is observed;
(P_2) But if A were true, C would be a matter of course.

(C) Hence, there is a reason to suspect that A is true."

In Peirce's original definition the selection and evaluation of explanation A is all part and parcel of the same inference. But usually not only the truth of A would make C a matter of course. Say B could also lead to the truth of C. So clearly Peirce's definition is not enough for an inference to the *best* explanation. Thagard made a clear distinction between the inference of possible explanations for surprising facts, and their evaluation. Peirce's original definition of abduction is a clear form of inferring from P_2 a possible explanation for P_1. But before jumping to conclusion C, other known premises like P_2 should be considered first.

Thagard defined a separate process to evaluate the resulting set of possible explanations, and called that process inference to the best explanation (IBE). Thagard defined IBE as an inference to a known explanation which explains the highest number of other known facts, needing the lowest number of auxiliary hypotheses as background assumptions. An explanation's value can be calculated by subtracting the number of auxiliary hypotheses from the number of explained facts. In that way, adding an explained fact *ad hoc* by an auxiliary hypothesis makes no difference for an explanation's value.

In PI, abduction and IBE are modeled as a process of problem solving. An explanation problem is represented by a basic memory structure, including the slot `START` containing context facts, and the slot `GOAL`, containing the explananda, the facts to be explained. Theories are represented as (secondary) processes called *rules*, with slots `CONDITION`, which contain premises and `ACTION`, containing conclusions. When a problem is set, a primary process of spreading activation activates rules linked to the problem slots. Only active rules are used to infer possible explanations for the slot value of `GOAL`. IBE decides which explanation is the most favorable. For example, we have three possible explanations of an observation E, represented in three rules. Activation from E activates the rules, which generate possible explanations by abduction. IBE selects the best as a conclusion of solving the explanation problem, see Table 5.2.

In PI, rules, problems and concepts all have basic structure types. Among the basic slots are `ACTIVATION`, `STRENGTH`, and `OLD-MATCHES`. The processes of activation, abduction and IBE are all primary. Only instances of concepts, rules and problems are secondary. IBE in PI is a process specially used for making evaluations of abductions, which only occur during explanation problems.

Explanation	Structure	Process	Example
	EXPLANATION		
Premise	START		F (is known to be true)
	GOAL		E (is to be explained)
Background		RULE-1	CONDITION H1 ACTION E
		RULE-2	CONDITION H2 H3 ACTION E
		RULE-3	CONDITION H4 ACTION E F
Inference		Activation	(E activates rules 1 to 3)
		Abduction	H1, H2&H3, H4 (possible explanations)
		IBE	H4 (explains the most facts with the least auxiliary hypotheses)
Conclusion			H4 (is the best explanation)

Table 5.2: Explanation as modeled in the PI program

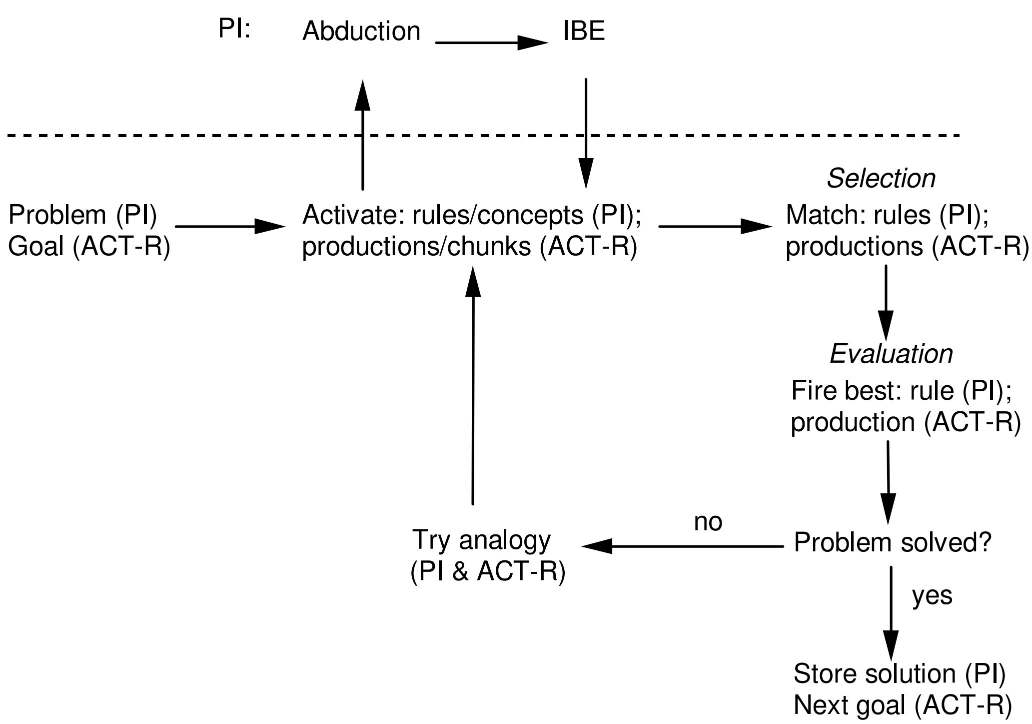

Figure 5.1: Problem solving in PI and ACT-R

Problem solving in ACT-R is similar to that of PI (see Figure 5.1), but with a few important differences. In both PI and ACT-R memory structures match with rules, which can add to memory and influence problem solving control. Yet productions in

ACT-R are of a different type than PI's rules. They represent a skill, and not an explanatory relation. And, more important for modeling explanation, ACT-R lacks a primary abduction mechanism. Because of the nature of the ACT-R theory such a primary mechanism is not appropriate. Productions in ACT-R are steps of practiced problem solving, generalized from example problem solutions by analogy (PI can also employ analogy to suggest rules, but I will not go into that here). So if a cognitive model in ACT-R needs to employ abduction in problem solving, then the abduction inference rule has to be learned first. And that turns out to be no problem at all.

Learning abduction by example - part 1

The ACT-R theory assumes that part of the process of solving a particular problem, is trying to recall an example of a problem that was solved earlier and had a goal similar to the current problem. When such an example problem is retrieved from memory, the structure of that example problem, and the solution of that example problem, is mapped to the current problem. When the solution of the example problem can be used to solve the current problem, a production rule is proposed, as a generalization of a strategy for solving problems that share the particular goal. It is currently assumed that all procedural skills, represented by production rules, are learned by this process of generalization from declarative examples.

The discussion and models in this section are based on the analogy mechanism of ACT-R, release 3.0. The details of implementing the mechanism of analogy have been changed in the 4.0 version that was introduced after I wrote this chapter. Learning by examples in ACT-R is studied extensively by Niels Taatgen (1999).

In this subsection I model an example of Paul Thagard's from his (Thagard 1988). He tells about his encounter with a group of outrageously dressed persons at the airport. He wonders why theses people are dressed up that way. Maybe they are rock musicians, he thinks, because rock musicians usually dress outrageously. ACT-R has to know only this example to generate, by analogy, a production that can make similar abductive inferences in the future.

As a similar explanation problem I use another example from (Thagard 1988). In this simple historical example the goal is to explain why sound propagates. It is known that waves propagate, so maybe sound is a wave. I started out with the following memory chunks:

```
(Example-Problem
    ISA                 explanation-problem
    GOAL                Dressed-Outrageously)
(Example-Rule
    ISA                 pi-rule
    CONDITION           Rock-Musician
    ACTION              Dressed-Outrageously)
(Example-Solution
    ISA                 explanation-solution
    EXPLANATION         Rock-Musician)

(Example-Dependency
    ISA                 dependency
    GOAL                Example-Problem
    SUBGOALS            Example-Solution
    CONSTRAINTS         (Example-rule))
```

```
(Problem-1
   ISA                 explanation-problem
   START               sound
   GOAL                Propagates)
(Rule-1
   ISA                 pi-rule
   CONDITION           wave
   ACTION              Propagates)
```

The example-dependency chunk is used (In AC-R 3.0) to represent the link between a problem chunk and the chunk that represents the solution to that problem. The constraint slot is used to represent that additional chunks that where involved in solving the problem.

The slot values `Rock-Musician`, `Dressed-outrageously`, `Propagates`, and `Wave` are also added as memory chunks of type `concept`. This chunk type also has a slot `INSTANCES`, which is filled with `Sound` for concept `Propagates`. The goal focus is set on `Problem-1`, which represents the problem to explain why sound propagates.

When ACT-R is started it first tries to match the goal `Problem-1` with available productions. After failing to do so (there are none defined) ACT-R searches for an analogous problem and finds `Example-Problem`. The special dependency chunk is used to find its solution. ACT-R uses the `Example-Rule` to map the solution to the problem, and uses it to make a new production. It then tests whether the new production will match the focused goal. Only if that succeeds will the new production be added to production memory. In my example ACT-R produces the following production:

```
(EXPLANATION-PROBLEM-PRODUCTION0
  =Example-Problem-Variable>
    ISA                        explanation-problem
    GOAL                       =dressed-outrageously-variable
  =Example-Rule-Variable>
    ISA                        rule
    CONDITION                  =rock-musician-variable
    ACTION                     =dressed-outrageously-variable
==>
  =Example-Solution-Variable>
    ISA                        explanation-solution
    EXPLANATION                =rock-musician-variable
  !focus-on! =Example-Solution-Variable)
```

The first condition chunk matches with `Problem-1` and the second with `Rule-1`. As a result the production creates a solution and changes focus of attention to it. This rule now serves as a secondary simple abduction process, generating hypothetical explanations, given explanation problems and rules that may explain it. The resulting explanation for the example is:

```
(**Example-Solution-Variable$1>
    ISA                        explanation-solution
    EXPLANATION                Wave)
```

This example has only one rule to abduce from. Usually several rules can be used to generate an explanation. Thagard employed IBE in PI to evaluate possible explanations before jumping to a best conclusion.

It can be argued that the general idea of Thagard's IBE is subsumed by ACT-R's primary, processes that subsymbolically select and decide which chunks and productions to match. Thagard's IBE favors the hypothesis that explains most known facts with the least number of auxiliary hypotheses. So there is a constraint on explanatory success and hypothesis simplicity. The simplicity constraint is met by ACT-R's primary process of latency, which is related to the probabilistic evaluation whether a chunk is relevant in a particular context. A more complex rule will contain more chunks in the condition, which will take longer to match. So more simple hypotheses will be considered first. Yet a very successful rule will have a higher activation because it is associated with more active facts in memory. So the constraint on explanatory success, is met by the process of preferring high activation.

One could compare the effect of the activation of chunks as a result of their probabilistic association with other chunks in ACT-R, with the effect of the activation of propositions as a result of their explanatory relation with other propositions in ECHO, Thagard's refined explanation evaluation model (Thagard, 1992).

Yet, several other factors, such as the production's expected gain (PG-C) value, play a role in the final decision to fire a rule. Hence ACT-R might not always come to similar conclusions as PI. Whether ACT-R's conclusions are more plausible is another question altogether, belonging to primary epistemology. However, because of the fact that ACT-R is a more sophisticated model of primary cognition than PI is, ACT-R is likelier to make abductive inferences that are closer to actual human problem solving. Whether that is relevant for epistemology is discussed in section 5.8.

Heuristic search as abduction in BACON

The BACON models (Langley et al, 1987) constitute a set of productions that try to find algebraic laws that describe given numerical observations. Several versions of BACON were originally implemented as a set of productions in the problem solving architecture PRISM, an old cousin of ACT-R (they both have ACTE as an ancestor). Hence it was relatively easy to model BACON.1 in ACT-R. Yet, a distinguishing claim of the ACT-R theory is that productions are not learned passively by *e.g.* reading, but by analogy during problem solving, by doing. Therefore I tried to model learning BACON's main productions by analogy. Doing so made apparent that in fact BACON's heuristic search method makes use of abductive inference in a way similar to PI's method.

The first of the BACON series searches for simple algebraic laws, which are all of the form $X^k Y^l = aX^m Y^n + b$. It tries to find appropriate values for *k, l, m, n, a* and *b* given a set of different observed values for X and Y. Laws that fit this template are, for example, Kepler's third law of planetary motion $D^3 P^2 = k$, Boyle's gas law $PV = c$, Galilei's law of acceleration $D/T^2 = g$, and Ohm's law $IL = -rI + v$.

BACON.1's search starts out with two observational terms X and Y, together with a set of values. For example, X is (1 2 4) and Y is (1 0,5 0,25), meaning that when X is 1 Y is 1, etc. The next step is to combine two terms as a product or a ratio and evaluate the resulting set of values, *e.g.* X*Y is (1 1 1). When the values of a term are found to be constant, a law is inferred. In the example $X*Y=c$. The same happens when two terms are related linearly. If the new term does not turn out to have constant values, or to be linearly related with other terms, then it can be used to make a next new term by combining it with the other available terms, *e.g.* (X*Y)*Y.

The BACON productions do not produce new terms at random, but *heuristically*. A heuristic method does not guarantee that a solution will be found, but often a solution can be found without evaluating every possible solution by brute force search. BACON.1's heuristic term generation is implemented in productions called Increasing and Decreasing. These productions determine what new term to consider as a possible law. Given that the absolute values of two terms both increase Increasing suggests to consider their ratio as a new term. Decreasing suggests to consider the product of two term when the absolute values of one terms decrease while the absolute values of the other increase.

These productions, together with the main productions that implement the search process are listed in Table 5.3. The search process itself is depicted in Figure 5.2, and summarized in Table 5.4. As an example, I listed the terms used and defined in the process of finding Kepler's third law of planetary motion in Table 5.5, based on Borelli's observations of the moons of Jupiter that were discovered by Galileo.

Production	Conditions (LHS)	Actions (RHS)
Find-Laws	Goal = describe data Law not already defined?	New goal = find-laws
Increasing	Goal = find-laws Term-1 increasing values? Term-2 increasing values?	New goal = consider-ratio
Decreasing	Goal = find-laws Term-1 increasing values? Term-2 decreasing values?	New goal = consider-product
Constant	Goal = find-laws Term constant values?	New goal = define-new-law
Linear	Goal = find-laws Term values linear related?	New goal = define-new-law
Define-Ratio-or-Product	Goal = consider-ratio/product Term not already defined?	New goal = define-new-term

Table 5.3: Overview of the main productions of BACON.1

ACT-R can learn the productions Increasing and Decreasing from given examples. The examples I used constituted algebraic rules that can be used abductively by ACT-R's process of analogy. For example it is true for the function X/Y=*c*, that if the absolute values of X increase, the absolute values of Y increase as well. On the other hand it is true for the function X*Y=*c*, that if the absolute values of X increase, the absolute values of Y decrease (see Figure 5.3).

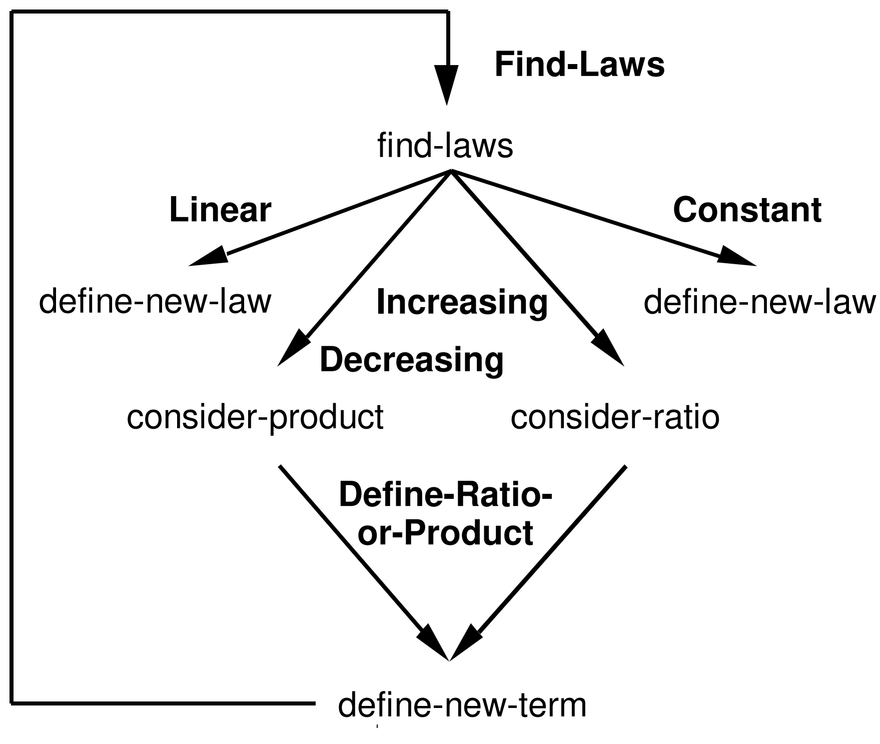

Figure 5.2: BACON.1's search for a law with constant or linearly related values

Description	Structure	Process	Example
Premise	X		1 4 9
	Y		1 8 27
	Goal		Describe X and Y
Background		Production-1	Find-Laws
		Production-2	Increasing
		Production-3	Decreasing
		Production-4	Constant
		Production-5	Linear
		Production-6	Define-Ratio-or-Product
Inference		Repeated matching of production rules	
Conclusion	Law X Y		$X^3/Y^2=1$

Table 5.4: Inferring a description in BACON.1

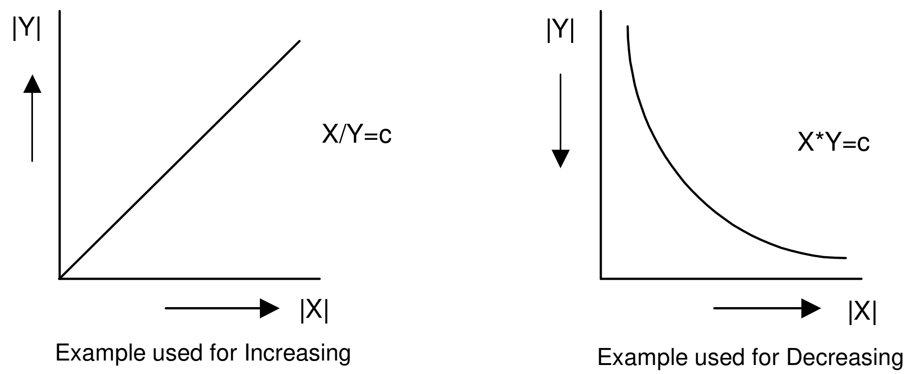

Figure 5.3: Example functions used for creating the main BACON productions

The other way does not always hold. For example, when both values of an X and Y increase, the relation may just as well be an exponential function. Hence BACON.1's productions actually infer by abduction that X and Y are related as a product or ratio. Increasing employs actually the following abductive inference:

The absolute values of X and Y both increase (C)
But if X/Y = c (A) then the absolute values of X and Y would increase (C)

Hence there is a reason to suspect that X/Y = *c* (A)

If the inferred new term is not evaluated to be a law, like *e.g.* D/P in the Kepler example, then values of the term can be treated as part of the background in a new abductive inference. The same compositional process is used in PI, see for example Table 5.6.

Explanation	Structure	Process	Example
Premise	PROBLEM		
	START		A (is known to be true)
	GOAL		C (is to be explained)
Background		RULE-1	CONDITION B ACTION C
		RULE-2	CONDITION A ACTION B
Inference		Activation	(C activates RULE-1)
		Abduction	B (possible explanation)
		Activation	(B activates RULE-2)
		Abduction	A (possible explanation)
		IBE	A
Conclusion	EXPLANATION		A (is the best explanation)

Table 5.6: Example of compositional abduction in PI

Learning abduction by example - part 2

To learn ACT-R BACON's heuristics I provide the functions of Figure 5.2. as solutions to a BACON search problem. The example for Decreasing was given as follows:

```
(X1>
  ISA                              term
  PATTERN                          Increasing
  EXP                              Example-Experiment)

(Y1>
  ISA                              term
  PATTERN                          Decreasing
  EXP                              Example-Experiment)

(Example-Problem1>
  ISA                              find-laws
  EXP                              Example-Exp
  ACHIEVED-BY                      Consider1)

(Example-Solution1>
  ISA                              consider
  OP                               Product
  TERM-1                           X1
  TERM-2                           Y1
  CONSTRAINTS                      (Decreasing Increasing))

(X1*Y1>
  ISA                              term
  VALUES                           nil
  OP                               Product
  TERM-2                           Y1
  TERM-1                           X1
  PATTERN                          Constant)
```

When the product of X and Y is constant, the values of X increase while the values of Y decrease. So if you want to find a law for the terms of an experiment Example-Exp, which are X1 and Y1, then by abduction BACON should consider their product. The production Consider-Ratio-or-Product would then define the term X1*Y1. To trigger ACT-R's analogy mechanism I set the following problem:

```
(Pressure>
  ISA                              term
  PATTERN                          Increasing
  EXP                              Boyle-Exp)

(Volume>
  ISA                              term
  PATTERN                          Decreasing
  EXP                              Boyle-Exp)

(**Boyle>
  ISA                              find-laws
  ACHIEVED-BY                      nil
  EXP                              Boyle-Exp)
```

By analogy with `Example-Problem-1`, using chunk `example-experiment` to map the solution to the problem, ACT-R composes a new production to solve the Boyle problem:

```
(FIND-LAWS-PRODUCTION1
  =Example-Problem-Variable>
    ISA                              find-laws
    EXP                              =example-exp-variable
  =Y1-Variable>
    ISA                              term
    EXP                              =example-exp-variable
    PATTERN                          decreasing
  =X1-Variable>
    ISA                              term
    EXP                              =example-exp-variable
    PATTERN                          increasing
==>
  =Example-Solution-Variable>
    ISA                              consider
    OP                               product
    TERM-1                           =X1-Variable
    TERM-2                           =Y1-Variable
  !focus-on! =Example-Solution-Variable)

(**Example-Solution$1>
    ISA                              consider
    OP                               product
    TERM-2                           Volume
    TERM-1                           Pressure)
```

The analogy mechanism of ACT-R (3.0) would overgeneralize the example problem without further constraints. The resulting production would match any two terms and consider their product. Yet with constraints, the inferred rule is functionally equivalent with BACON's original Decreasing production, and can hence be employed to find more complex laws.

In sum, the computer programs BACON and PI model cognitive mechanisms of scientists that work on particular scientific problems. In this section I argued and showed how those same mechanisms can be learned and explained, by modeling that learning process in the unified cognitive theory ACT-R. Yet between the different cognitive architectures there remains a difference in approach to understanding the nature of scientific theory and reasoning. This is treated in the next section.

5.6 Theory and method

In this section I go further into the specific questions about the structure of theory and process of reasoning as implied by the different cognitive models BACON, PI, and ACT-R. Thagard's model PI maintains a procedural explanation of the nature of a theory. In the logical approach a theory is a set of atomic and conditional propositions, accompanied by a set of relatively independent inference rules that are used to infer valid consequences from them.

In Thagard's model a theory consists of rules and concepts that more or less represent conditional propositions and predicates, respectively. Which inference rule to apply to determine a consequence of a theory is arbitrary in logic, but controlled by a mechanism of spreading activation in PI. Superficially, this difference only has consequences in the performance of the process of generating an explanation or prediction. In principle the same consequences could be inferred from the different representations of a theory in both the cognitive and the logical model.

Even in the process of inferring an explanation the main difference between the cognitive model PI and the logical model lies in their performance and the specific extra conditions. Thagard selects the best explanation on the equally decisive criteria of explanatory breath and simplicity, while the logic approach puts the priority on explanatory breath and consistency.

One important difference is that PI maintains different theories simultaneously, basing the use of any of the rules in prediction or explanation on its success in solving problems earlier. This allows PI's predictions to be inconsistent due to the firing of competing rules.

The nature of the heuristic rules of BACON differ in type from those of PI. The BACON heuristics represent a very specific kind of abduction. PI's primary abductive mechanism reasons from all kinds of conditional assumptions represented in the PI rules. The BACON heuristics incorporate an abductive suggestion based on a conditional proposition, *e.g.* the proposition that if the quotient of two variables is constant, then the values increase together. Any term proposed by those heuristics (INCREASING/DECREASING) is tested on the available data terms by other heuristic rules (LINEAR/CONSTANT) that propose it as a law or ignore it if it does not fit the data.

The BACON production rules implement a particular heuristic method, and not a part of a theory as in PI. The rule representation of either a theory or heuristic is subtle. For the predictive nature of a theory it is not important whether you represent a theory as a set of conditional statements or as a set of production rules, as long as the specific production rules, or the conditional statement together with general inference rules produce or define the same consequences.

It is possible to understand a theory both declaratively and procedurally in ACT-R. The structure of a theory can start out as a declaration in memory chunks. What consequences will follow from it depend on the production rules that can make an inference about it. It is possible to represent both the axioms of a theory and general inference rules declaratively, and a method to infer deductive consequences from them can be represented by a set of productions.

I summarize the different uses of the rule concept in explanation models in Table 5.6. Rules can be considered to be secondary processes in all the cognitive models. But in PI they are part of theory, while in BACON.1 they are part of method. In ACT-R a production can be understood as both part of theory and/or heuristic method.

In ACT-R a production, seen as an inference procedure, takes as premises a goal and an assumption, and produces a new goal as conclusion. This goal can be either to make a new assumption, to observe or to intervene within something in the world. If the premise of a specific production includes declarative assumptions of concept types A and A $\rightarrow$ C and the goal is to produce a valid consequence, then the production represents the application of *modus ponens* if its new goal is to assume C.

Explanation	Logic	PI	BACON.1	ACT-R
Background Structures	B: {A→C}		(if X/Y=c then inc)	Chunk: (rule A C)
Processes		(If A then C) Rule = theory	(If inc then ratio) Rule = method	(If (rule A C) goal C then A) Rule = method & theory
Premise	P: {C}	(START C)	(goal) (X …) (Y…)	(goal C)
Inference	Abduction Conditions	Activation Abduction IBE	Rule matching	Creation (Analogy) Selection (Activation) Evaluation (PG-C)
Conclusion	H*: {A}	(EXPL. A)	(X^nY^m=c)	(A)

Table 5.6: Different uses of the rule concept in explanation

But how to understand the generation of specific explanations? As we saw in the earlier sections, in ACT-R this question is not so much about what productions can find an explanation, but how productions that can find explanations are created and evaluated themselves. This is a process in ACT-R that starts with an example of a specific explanation and a similar example solution to another problem. The example is mapped to the new problem, resulting in new productions. These productions can become either applicable to very specific cases or very general cases, for which the inferred explanation has a very high or low probability of being correct. By solving many explanation and prediction problems, use and experience will determine their success. The resulting productions can be associated with the typology of strong and weak heuristics, see Table 5.7.

The term heuristic comes from the Greek *heuriskein* meaning “to discover”. (*Heuriskein* is also at the origin of *eureka*, derived from Archimedes’ reputed exclamation, *heurika* (for “I have found”), uttered when he had discovered a method for determining the purity of gold by taking a bath) In artificial intelligence it is generally used to describe a process of learning by trying. It is often contrasted by the term algorithm, which is a derivation of the name of the Arab mathematician, Al-Khowarizmi (±825 AD). Both an algorithm and a heuristic are procedures for solving a problem.

The main difference between them is that an algorithm is meant to effectively solve a particular type of problem, often at high cost in time depending on the complexity of the problem. A heuristic is a tradeoff between time and optimality, it may solve a problem, usually at lower cost in time, but then it may not provide the best solution. Another difference is that the effectivity of an algorithmic procedure can usually be established analytically by mathematical proof, while the effect of a heuristic procedure is often established empirically, by experience of use.

Productions	High cost	Low cost (efficient)
High probability (effective)	Strong heuristic Specific method/theory	…
Low probability	…	Weak heuristic General method/theory

Table 5.7: Typology of productions in the light of expected gain (E = PG – C)

The heuristic procedure in ACT-R differs from the static heuristic procedure in BACON in such a way that the estimation of the cost and chance of success of a certain production (the estimated gain PG – C) is constantly evaluated and adjusted. So if we want to explain the process of discovering a theory or law it is not enough to point to a set of heuristics as the cause of that discovery. The heuristics are usually part of the product of that discovery. For Kepler to discover his law he first had to discover that he could compare Borelli's data with a particular kind of example functions.

Now if we understand the nature of a theory as being partly procedural we can also better understand how to see Kuhn's picture of science as a practice that reasons on the basis of paradigms as shared examples. In normal science a set of successful examples of explanations leads to a strong heuristic that can successfully solve highly specific problems within a domain. At a given time these heuristics, that incorporate part of the theory of that domain, may not be able to handle novel problems. A revolution is needed to start off a different approach, where only weak heuristics may be of some help. More specific and stronger heuristics will be learned once some success is booked.

So, to understand a theory and use it rationally is to learn a skill of a specific practice. You can tell a lay person that $E=mc^2$, but without a general skill in mathematics and a specific skill of how to apply those variables to a domain of phenomena, that person will not be able to predict or explain specific facts with that statement.

In Kepler's and Galileo's time science had become successful by applying simple and general mathematical functions to empirical phenomena en testing the predictions of those functions in experiments. But the practice of empirical science consist of the use of many, highly specific, constantly adjusted rules to explain and predict phenomena. A reflection of those rules can be declaratively represented and communicated, that is what this thesis is all about. Their use cannot be learned otherwise then by taking part in that practice. But is the way scientists actually use method and theory a criterion for what is rational, from an epistemological point of view?

5.7 Descriptive and normative

I argued how the ACT-R theory can provide an explanation of the rational behavior of scientists. But what does that tell us about what is rational? It is argued in epistemology that explaining the beliefs and methods of scientists by pointing to the cognitive process that creates and evaluates them is not sufficient for epistemically justifying those beliefs and methods.

People, scientists not excluded, make mistakes in their reasoning, as psychological experiments prove. It should be the role of epistemology to point out those errors, so that human reasoning can improve. So let us look at the rationality of human prediction and explanation.

Prediction

Human performance in logical reasoning has been a much studied subject in cognitive psychology (Anderson, 1995). In experiments by *e.g.* (Marcus & Rips, 1977) subjects were asked to evaluate the correctness of hypothetical syllogisms, represented as relatively neutral arguments such as:

> If the ball rolls left, the lamp will switch on.
> The ball rolls left
> Therefore, the lamp will switch on.

It was asked if a conclusion is always, sometimes, or never correct. It was shown that in 100 percent of the cases subjects have no problems with judging the conclusion of *modus ponens* (affirming the antecedent) to be always correct, but that in only some 80 percent of the cases subjects judged the conclusion of denial of the antecedent and affirming the consequent to be merely sometimes correct. Still worse, in only 60 percent of the cases subjects thought that *modus tollens* (denying the consequent) is always correct. This performance was initially explained by the assumption that subjects interpret “if A then C as a biconditional instead of a conditional statement. Subjects were thought to understand the antecedent to be a necessary condition for the consequent, explaining why in some cases it was thought that the conclusion of denial of the antecedent or affirming the consequent is always correct. However, this does not explain the poor performance on judging the validity of *modus tollens*.

It is remarkable to see that the inference that is the hallmark of valid reasoning in science according to Popper is so often misjudged in common sense reasoning. It also testifies to the unpopularity of Popper’s method of falsification as noted by Kuhn and many others. But it would be too swift a conclusion to mark the disregard of *modus tollens* in the practice of both common sense and scientific reasoning as irrational. It becomes more clear if this practice is seen as based on a probability assessment. I will demonstrate this by discussing another much studied task, called the Wason selection task.

Understanding the performance of subjects on the selection task is relevant for understanding how scientists evaluate potential hypothesis in the process of scientific discovery. This task is argued to demonstrate the failure of applying *modus tollens*. However, I will argue how this task shows how subjects make a perfectly rational probabilistic assessment.

In the selection task subjects are shown four cards with the following symbols:

E	K	4	7

They are told that every card contains a number on one side and a letter on the other.

The task is to test the validity of the following rule for these four cards:

If there is a vowel on one side, then there is an even number on the other side.

Subjects were asked to turn over only those cards that need to be turned over to test the rule. On average (Anderson, 1995) 89 percent chose to turn the E, affirming the antecedent of the rule. Logically this is an informative choice because the outcome of the experiment either falsifies or confirms the rule. However, 62 percent chose to also turn over the 4, affirming the consequent. Logically this provides no information because the outcome confirms the rule either way. The same goes for turning over the K denying the antecedent, which was done by 16 percent. Only 25 percent chose to turn the 7, denying the consequent, which logically also can confirm or refute the rule.

Oaksford and Chater (1996) argued that what subjects do is make a choice of the most informative cards in a statistical sense. They presupposed a probabilistic model of the rule A $\rightarrow$ C, see Table 5.8. It provides the probabilities for the four possible states of the world where A and C are either true or false. Given this probabilistic model for the rule A $\rightarrow$ C and a null rule, *i.e.* a rule which does not have any probabilistic contingency between A and C, the interpretation of the conditional probabilities of A and C can be calculated, see Table 5.9a, see also Table 4.4 prediction.

Given the probabilistic interpretation both the AA and DC predictions are probable, while AC and DA are less probable. Yet subjects prefer AC much more than DC. To explain this, Oaksford an Chater argued that a card would be informative if the expectation of its outcome would differ from the expectation based on a null rule that assumes no relation between the antecedent and the consequent. However, in their model they need to set the conditional probability of the consequent C, given the antecedent A and vice versa to be 40% instead of a neutral 50% to explain the preference order of subjects.

Antecedent A	Consequent C	A $\rightarrow$ C	A $\rightarrow$ C	null	A $\rightarrow$ C *	null *
True	True	True	.40	.16	.50	.25
False	True	True	.20	.24	.23	.25
False	False	True	.30	.24	.18	.25
True	False	False	.10	.36	.09	.25

Table 5.8: The logical, and two possible probabilistic models of A $\rightarrow$ C and null

I think that there are three problems with the explanation of Oaksford and Chater. First, the particular probability distribution of the conditional statement is not properly defended. Secondly, a proper 50/50 null rule defeats their ordering. Thirdly and most importantly, the probability of a rule's prediction does not reflect the rule's probability given the outcome of the experiment.

It may well be possible that for subjects the probability of a rule A $\rightarrow$ C depends on the assumed model of the rule, not on the probability of a rule's prediction. In this interpretation the value of an experiment is the difference between the probabilities of a rule given the possible outcomes of that experiment. Given this interpretation the second problem becomes obsolete by addressing the first problem.

What is a proper model for a general conditional statement? One could argue that the preference of subjects in the card selection task actually reflects an average model for a conditional rule. If we redistribute the preferences of subject over 100% and take that as a value estimate, then we come to an average model that is approximated in Table 5.8 for rule A → C *. In this estimate subjects tend to regard the average probability of a rule slightly higher when only C is observed (.23), compared to when A nor C is observed (.18), see Table 5.9 b. It can be assumed that these numbers at best reflect a base rate probability that is different and adjusted for every particular conditional assumption that is maintained in memory.

	B	H	P	p(P|B&H)	p(P|B & null)	Difference	Subj. pref.
a.							
AAH	A	A → C	C	.80	.40	.40	89% E
ACH	C	A → C	A	.67	.40	.27	62% 4
DCH	Not C	A → C	Not A	.75	.60	.15	25% 7
DAH	Not A	A → C	Not C	.60	.60	.00	16% K
b.					p(P|B & null*)		
AAH	A	A → C *	C	.84	.50	.34	89% (47%)
ACH	C	A → C *	A	.68	.50	.18	62% (32%)
DCH	Not C	A → C *	Not A	.67	.50	.17	25% (13%)
DAH	Not A	A → C *	Not C	.56	.50	.06	16% (8%)
							(100%)
c.				p(H|B&P)	p(H|B & ¬P)		
AAH	A	A → C *	C	.50 (C)	.09 (R)	.41	47%
ACH	C	A → C *	A	.50 (C)	.23 (C)	.27 (-.14)	32% (-15)
DCH	Not C	A → C *	Not A	.18 (C)	.09 (R)	.09 (-.18)	13% (-19)
DAH	Not A	A → C *	Not C	.23 (C)	.18 (C)	.05 (-.04)	8% (-5)
d.							
HAA	A → C *	A	C	.68 (?)	.33 (R)	.35	
HAC	A → C *	C	A	.84 (C)	.44 (?)	.40	
HDC	A → C *	Not C	Not A	.56 (?)	.16 (R)	.40	
HDA	A → C *	Not A	Not C	.67 (C)	.32 (?)	.35	

Table 5.9: Different kinds and models of probabilistic prediction

So given the above model the value of AA is the highest because that model assumes the rule is the most probable if A and C are true (.50) and the least probable if A is true and C is false (.09). The value for DA is the lowest because either outcome says about the same (.18/.23) about the probability of the rule, given the model. The reason that AC is more preferable than DC is that the difference between the outcomes for the former experiment is much higher (.50 – .23) than that of the second (.18 – .09). The outcome of DC may logically be able to either defeat or confirm the rule given the logical model of the rule, but with a probabilistic model either outcome of a DC experiment will result in a low probability.

To make the comparison with the logical model complete I also listed the kind of predictions where the rule is assumed and the antecedent or consequent is hypothetically affirmed or denied. A probabilistic interpretation now provides an assessment where the logical approach could not give an answer about the probability of the hypothesis, see Table 5.9d.

From this viewpoint subjects predictions and experiments do not seem to be all that irrational, as long as hypotheses are interpreted to be more or less probable instead of just true or false. In a game like situation, where the rules are strict and given, it is rational to follow the logical model of a rule. But in an empirical situation where rules are not known to be true and almost all rules have exceptions acting on a probabilistic assessment is more rational. Yet Popper would probably argue that the question remains how probability assignments to hypotheses can be rational. This question will be addressed in the Chapter 6.

5.8 Explanation and evaluation

According to Langley et al (1987, p.47) in discovering a hypothesis "rationality for a scientist consists in using the best heuristics available for narrowing the search down to manageable proportions. A normative theory of creativity and scientific discovery is concerned with this kind of rationality." So instead of focussing on the validity or probability of hypotheses found by heuristics, they emphasize the efficiency part of rationality. They assume you know what you are looking for. For Bechtel (1988) to normatively evaluate a heuristic is to identify its failure. He assumes you know when a heuristic fails. But how to know what you are looking for and how to know you failed to find it?

In epistemology it is a much debated question whether the identification of the failure or success of assumptions is an analytical or empirical matter. This holds for both theoretical and methodological assumptions. In a psychological explanation of scientific practice the identification of epistemic success or failure seems foremost an empirical matter. Productions and chunks are created and evaluated by their success in use, whether they are part of theory or of method. But the success of productions can only be measured by given conditions for success. And testing if a proposed solution satisfies those conditions is an analytical matter.

According to logic the best theory should be: consistent, internally and with respect to background knowledge; complete and correct with respect to the phenomena it explains; non-trivial; informative, and it should be simple. So different methods are suggested that prefer theories with regard to their competitors by their consistency, correctness and completeness, non-triviality, empirical content, and simplicity. Different philosophers prefer one condition above another on the basis of different arguments.

Scientists usually also entertain other preferences such as analogy, beauty or symmetry in a theory. Finding a theory that satisfies those conditions means success. But the most important condition of any theory is that it should remain successful in the future. So the questions with respect to the probability part of the rationality of reasoning are: 1. which conditions are conductive to empirical success, 2. why are they conductive to success, and 3. how to pursue them?

First there should be made a distinction between conditions that are part of the main goal of science, and those that may be conductive to it. I gather that conditions such as:

C_1. Correctness C_2. Consistency C_3. Completeness

are part of the main goal of science. This is what we want to achieve: a theory that has no anomalies, covers the domain, and is not trivial by allowing everything. These conditions are not conductive to empirical success, they define it. But how to pursue them? The satisfaction of the first two conditions can never be validly established in an empirical domain. The future can always bring a situation that is not allowed or included in the theory. The best we can do is to analytically check for internal consistency and to check for correctness and completeness with respect to all available observations. However: in principle, infinitely many theories can be entertained that satisfy all three conditions; in practice, however, it is hard to find even one theory that comes close to that goal.

Scientists build theories incrementally, constantly proposing and revising hypotheses, often within the conceptual boundaries of a research program. The question is whether it is rational to pursue correctness and completeness by preferring to pursue only a proposal that is closest to the goal. At any given time that goal seems clear. There is a set of current observations and the problem is to find that theory that covers most of them.

So, is it rational to entertain and pursue a consistent theory that explains most data and has the fewest number of counterexamples? By definition that theory is closest to the goal of science, assuming that all other possible theories are known to be worse. Yet in practice we do not know the merits of all possible other theories since we do not know them all. It may turn out that amending a theory that was further from the goal proves more successful than working on the best one available. Given a conceptual space of all possible theories and a set of all observations, the theory that best satisfies the goal at any moment of development may be stuck in a so-called local maximum. Pursuing predictions and revisions of a theory that is further from the goal may reveal a better approximation. In cognitive psychology and AI the first approach is know as hill-climbing. Going straight for the top may bring you to the top of the hill, but may miss the mountain. A scientist that chooses to stay with a successful theory that lacks progression is as rational as a chicken that gets stuck in a fence when running toward the corn in view, not able to back up to go around the open gate door.

In practice it does not always work that way. Scientists do not only pursue correctness, completeness and consistency. They also entertain conditions such as *e.g.*:

C_4. Simplicity C_5. Analogy C_6. Symmetry

In logic these conditions are meta-epistemical, they do not inform us about the truth of a theory. However, in scientific practice these conditions often prevail above correctness, completeness and consistency. (We will see an example of this in the case study in the next part of the thesis.)

Thagard incorporated these conditions in his theory of explanatory coherence, which meant to explain scientists' preferences. The program ECHO implements a model of a neural network that can evaluate how close a theory is to all conditions, as compared with a competitor (Thagard, 1992). Yet this theory fails to explain why it is sometimes rational to prefer conditions C_4-C_6 above C_1-C_3. How can these conditions, or methods based on them be conductive to the empirical success of a theory?

A naturalistic way out to this question is to explain why scientists have certain preferences by bringing in evolution, both biologically and socially. Primary mechanisms in our brain have preferences for certain assumptions and methods given experience. Survival depends on being able to make methodological decisions and retrieve memories of experiences that are relevant to the current situation or problem. An organism that is not able to make decisions or assumptions successfully is less likely to survive. In the development of our species nature favors particular primary cognitive mechanisms in the face of lions and gathering food; in the development of science nature favors particular theories, methods and scientists, in the face of peers and trying to get tenured positions.

To return to Goldman's distinction (Section 5.2): we have gone through an exposition of some (secondary) methods and theories and how they are generated and evaluated by some (primary) mechanisms of the brain during scientific discovery. I argued how these mechanisms tell us something about rationality. They inform us what rationality is, for a scientist.

However, these primary mechanisms still do not inform us *why* it is epistemically rational to maintain certain theories and methodologies. These mechanisms prefer a theory or method if it proves successful in solving problems, in reaching certain goals, satisfying certain conditions. But why are some conditions more rational to pursue than others, why are they more successful? A naturalistic stance would be happy with just the observation that certain conditions, methods, hypotheses and theories are more successful than others, as an inductively assumed fact of the world. Yet, in the next chapter I will pursue an explanation of one of those facts, why one of those conditions, simplicity, is conductive to attaining the goal of science.

Epistemologists reason to study reason is to be able to improve it. In this chapter we have come to understand reasoning as a process of inferring conclusions that satisfy certain conditions, given a certain problem. So to understand and evaluate the reasoning in a specific discovery process normatively it is first of all important to understand the details of a specific problem, *i.e.*:

- starting situation
- background assumptions
- process to reach the goal
- goal properties
- end results

In practice none of the above stay constant in the process. The starting situation changes, new background assumptions and concepts are added or withdrawn, end results are different from the goal, the goal conditions shift, and new methods to reach the goal are introduced. All under influence of primary cognitive mechanisms and social interaction. How this process goes about in the practice of neuropharmacology will be discussed in the next part of this thesis.

5.9 Conclusion

The particular question of this chapter was: how to understand and model scientific discovery, in ACT-R? I will answer this question by going through the answers for the specific questions of this thesis from Section 1.3:

Question 1 What is the structure of a scientific theory? In ACT-R theories can be understood as a collection of statements containing laws, examples and solutions to earlier explanation and prediction problems, represented declaratively in memory chunks, and specific and general procedures, represented in production rules. Chunks, represented as sets of slots and values of a certain type, are assumed to be the results of perception and solutions to solved problems. Production rules are represented as condition-action pairs: given a goal and an assumption chunk a new goal is set which can lead to either a new assumption or doing a particular observation or intervention in the world. Productions can be part of both theory and method.

Question 2 What is the process of scientific reasoning? The process of scientific reasoning in ACT-R contains of learning heuristic problem solving skills in searching and evaluating explanations and predictions of phenomena, see Table 5.10.

Problem	Start	Background	Process	Goal	Goal properties
Explanation	Goal = explain	H' explains P'	Creation	H*	H* explains P
	observation P	Productions	Selection		Analogy
			Evaluation		Probability
Prediction	Goal = predict	H' predicts P'	Creation	P*	H explains P*
	hypothesis H	Productions	Selection		Analogy
			Evaluation		Probability

Table 5.10: Short overview of reasoning problems discussed in this chapter

The process of both explanation and prediction starts with a goal chunk together with examples and productions in memory in the background. A solution to the problem is either selected from memory by productions or created based upon examples by analogy, and evaluated probabilistically.

Question 3 What is the route between theory and experiment? The assumed route between theory and experiment walked by a scientist starts with a goal and assumptions in memory that determine new assumptions and actions, based on learned productions. Failure to achieve a goal decreases the potential to recall an assumption and the chance that the used productions will be employed in the future.

This can explain how scientists go through the ideal six steps introduced in the last chapter. In a scientific study of scientists doing their work you would get the following scheme, where lowercase p denotes a phenomenon and uppercase P a proposition about that phenomenon:

1. Observe phenomenon p: see $p_m,\ldots,p_n$ (activities of scientist x at work)
2. Describe p: $P_m \rightarrow P_n$ (problem solving behavior)
   P_1: {x observes phenomenon *p*: x sees p_m}
   $P_1 \rightarrow P_2$: {x describes *p*: $P_m \rightarrow P_n$}
   $P_2 \rightarrow P_3$: {x explains *p*: x finds $B \cup H^* \models P_m \rightarrow P_n$}
   $P_3 \rightarrow P_4$: {x predicts *p*: x finds $B \cup H \models P_i^* \rightarrow P_j^*$}
   $P_4 \rightarrow P_5$: {x intervenes *p*: x creates P_i^*}
   $P_5 \rightarrow P_6$: {x observes *p*: x sees P_j^*}
3. Explain p: $B \cup H^* \models P_m \rightarrow P_n$
   H^*: {ACT-R cognitive mechanisms} $\models P_m \rightarrow P_n$
4. Predict p: $B \cup H \models P_i^* \rightarrow P_j^*$
   B: {specific chunks and productions of BACON}
   P_2^*: {x describes p: P_1 : $\{D = \langle 1, 4, 9\rangle\} \rightarrow P_2$: $\{P = \langle 1, 8, 27\rangle\}$}
   $P_2^* \rightarrow P_3^*$: {x explains p: x finds H: $\{D^3/P^2 = c\} \models P_1 \rightarrow P_2$}
5. Intervene in p: create p_i^*
6. Observe p: observe p_j^* ?

In this way a step in the process of scientific problem solving is described as a conditional statement. In step 1. the activities of a scientist are observed as a phenomenon. In step 2. these activities are described. One can observe a scientist making observations (P_1) and describing them (P_2). Logically one can describe the link between those activities by a conditional statement ($P_1 \rightarrow P_2$). The antecedent of the conditional statement represents the start situation, the consequent represents a goal situation. In step 2. of describing the activities of a scientist, one can further describe how a scientist explains (P_3) predicts (P_4) intervenes in (P_5) and again observes (P_6) a phenomenon. In step 3. of our cognitive research of scientific activities an hypothesis is searched to explain the process of those scientific activities, in this example cognitive models in the ACT-R architecture are proposed. In step 4. we make a prediction about how our scientist under study can find a law (P_3^*) that can imply data that describes a phenomenon (P_2^*). This prediction can be tested in step 5. and 6.

It can be a task for cognitive psychology to explain and predict how scientists search for a solution of scientific problems. For naturalistic epistemology it is the task to find an intervention in step 5. such that scientists can observe, describe, explain, predict and intervene the phenomena they are interested in more effectively and efficiently, and to explain why they do so. Why some explanations might be more effective than others will be the topic of the next chapter.

* * * * *

Chapter 6

Computation

6.1 Introduction

In the last two chapters we saw that both the logical and the cognitive models of scientific discovery include a condition to prefer simple or minimal explanations. None of the models further suggest why it is rational to prefer simplicity. I argued how the ACT-R model of cognition implicitly prefers simplicity as a consequence of a mechanism that prefers high probability in section 5.6 (page 67).

In this chapter I investigate the relation between probability and simplicity in the computational description, explanation and prediction of empirical data. I discuss the use of Kolmogorov complexity and Bayes' theorem in Solomonoff's inductive method to explicate a general concept of simplicity that is used for a distribution of probabilities of possible hypotheses. This makes it possible to understand how the search for simple, *i.e.*, short, descriptions of empirical data leads to the discovery of patterns in the data, and hence more probable predictions. I show how the simplicity bias of Langley's BACON.2 and Thagard's PI is subsumed by Rissanen's Minimum Description Length principle, which is a computable approximation of Solomonoff's uncomputable inductive method. A more lengthy discussion, including several other approaches to simplicity, can be found in (van den Bosch 1994).

In this chapter I pursue an answer to two particular questions: 1) How can simplicity most generally be defined? 2) Why should we prefer a simpler theory to a more complex one? I discuss simplicity definitions that stem from research in cognitive science and machine learning. In those approaches simplicity plays an important role in the process of scientific discovery, as implemented in Langley and Simon's computer model BACON, in inference to the best explanation, as implemented in Thagard's computer model PI, and in the probability of predictions, as explicated by Solomonoff.

Langley and Simon claim that the BACON programs search for simple consistent laws without making explicit what is meant by 'simple' laws and why we should pursue simplicity (Langley et al 1987). Thagard proposed an explicit definition of simplicity and employs it in his model PI, without providing a satisfying reason for it (Thagard 1988). However, Solomonoff proposed an explication of induction which makes use of a concept that can be used to understand simplicity and to provide a satisfying justification for its preference.

According to Solomonoff we should trust the theory yielding implications that can be generated by the shortest computer program that can generate a description of our known observational data. It is argued that a shorter computer program provides more probable predictions because it uses more patterns from that data. It is proved that this simplicity measure is reasonably independent of the computer language that is used. However, this measure has one drawback, it is uncomputable. Yet it is claimed that computable approaches to induction in machine learning constitute approximations of Solomonoff's method (Li and Vitányi 1994).

In this chapter I demonstrate how Solomonoff's approach can elegantly be used to make a universal prior probability distribution for Bayes' theorem. First it is shown that Rissanen's Minimum Description Length principle (MDL) can be derived from Solomonoff's approach. And from thereon I show that simplicity in Langley's BACON.2, and simplicity in Thagard's PI are nicely subsumed by MDL. I conclude this chapter by answering the three specific questions of this thesis, according to the study of computational description.

6.2 Turing machines

In 1964 an article by Solomonoff was published that contained a proposal for a general theory of induction. The objective was the extrapolation of a long sequence of symbols by finding the probability that a sequence is followed by one of a number of given symbols. It was Solomonoff's conviction that all forms of induction could be expressed in this form.

He argued that any method of extrapolation can only work if the sequence is very long and that all the information for an extrapolation is in the sequence. Solomonoff proposed a general solution that involved Bayes' theorem. This theorem requires that a prior probability of a hypothesis is known to determine the posterior probability making use of the known data. Solomonoff's solution is to provide for a universal distribution of prior probabilities, making use of a formal definition of computation.

It is widely believed that the notion of computation is fundamentally captured by the operation of a Turing machine, an idealized conceptual machine introduced by the mathematician Alan Turing. A Turing machine is thought of as consisting of a long tape and a finite automaton which controls a 'head' that can read, delete and print symbols on the tape. To compute the value of a function $y = f(x)$, write a program for $f(x)$ on the tape, together with a symbolic representation of x and start the Turing machine. The program is completed when the Turing-machine halts and the value of y is left on the tape as output. Turing proved that there is a universal Turing-machine that can compute every function that can be computed by any Turing machine. The famous Church-Turing thesis claims that every function that can be computed, can be computed by a Turing machine.

What Solomonoff did was to correlate all possible sequences of symbols with programs for a universal Turing machine that has a program as input and the sequence as output. He assigned a high prior probability to a sequence that can be computed with short and/or numerous programs. Sequences that need long programs and can only be compared by few programs, receive a low prior probability. For Solomonoff the validity of giving sequences calculated by a shorter program a higher prior probability is suggested by a conceptual interpretation of Ockham's razor. But it is justified because a shorter program utilizes more patterns in the sequence to make the program shorter. So, if we trust the data as being representative of things to come, then the shorter program provides the more probable predictions.

If we, *e.g.*, have a sequence that has x as a prefix and x = 1234123412341234123, then we could write a program that describes x as dαααα123, where d is a definition of 1234 as α. If we want to predict the following letter we can entertain the hypothesis H_1: dαααααα, or still shorter H_1: d5α. Another option is H_2: d4α1231. The first hypothesis predicts a 4 and the second a 1. Both hypotheses are compatible with the known data x. Now Solomonoff argues that the prediction of H_1 is more probable because it requires a shorter program to generate a continuation of x than H_2 (Solomonoff 1964, p.10).

That a sequence with many programs gets a high prior probability is suggested by the idea that if an occurrence has many possible causes, then it is more likely. The principle of indifference is integrated by attributing sequences that are generated by programs of the same length the same prior probability.

Unfortunately this approach has as an important problem. It is not determinable whether a given program is the shortest program that computes a sequence. If that were determinable then there would exist a Turing machine that could determine for every possible program whether it would generate a given sequence. However, most of the possible Turing machine programs will never halt. Due to this halting problem we cannot know for every program whether the program computes a given sequence. But before I go into that problem I want to make Solomonoff's theory more specific by first discussing Kolmogorov complexity and its application in probability theory.

6.3 Kolmogorov complexity

The Kolmogorov complexity of a sequence or string is actually a measure of randomness or, when inverted, the regularity of the patterns in that string. We can use a Turing machine to measure that regularity with the length of the shortest program for that Turing machine that has the sequence as output. We can call such a program a description of the sequence. This description is relative to the Turing machine that has the description as input.

So when we have a sequence x and a description program p and a Turing machine T we can define the descriptional complexity of x, relative to T as follows (*cf.* Li and P.M.B. Vitányi 1993, pp.352):

Definition 1 *Descriptional complexity*. The descriptional complexity C_T of x, relative to Turing machine T is defined by:

$$C_T(x) = \min\{ l(p): p \in \{0,1\}^*, T(p) = x \}$$

or $C_T(x) = \infty$ if no such p exists.

We consider T to be a Turing machine that takes as input program a binary string of zeros and ones, so the program is an element of the set $\{0,1\}^*$, which is the set of all finite binary strings. We use binary strings because everything that can be decoded, like *e.g.*, scientific data, can be coded by a string of zeros and ones. The length of the program, l(p), is the number of zeros and ones. So the definition takes as the complexity of a string x the length of the program p that consists of the least number of bits and that will generate x when given to T. If no such program exists then the complexity is considered to be infinite.

When x is a finite string then there is always a program that will describe it. Just take a program that will merely print the number literally. This program will be larger than the string. However, if x is infinite and no finite program exists, then x is uncomputable by definition.

This complexity measure is relative to but surprisingly largely independent of the Turing machine in question, as long as it is a universal Turing machine. There exists a universal Turing machine that computes the same or a lower complexity than the complexity computed by any other Turing machine plus some constant dependent on that other Turing machine. For instance, when I have a string and two programs in different computer languages that compute that string, the difference in length between those programs cannot be more than a constant, independent of the string. This claim is called the invariance theorem (*cf.* Li and Vitányi 1993, pp.353).

In the literature Kolmogorov complexity K(x) is defined as a variant of descriptional complexity C(x), which makes use of a slightly different kind of Turing machines. In the definition of descriptional complexity a Turing machine was used with one infinite tape that can move in two directions and that starts with an input program on it and halts with a string on the tape as output. For the definition of Kolmogorov complexity a prefix machine is used. This kind of Turing machine uses three tapes, an input tape and an output tape which both move in only one direction, and a working tape that moves in two directions. The prefix Turing machine reads program p from the input tape, and writes string x on the output tape.

Kolmogorov complexity will render a similar measure of complexity as descriptional complexity, where C(x) and K(x) differ by at most 2 log K(x). This difference is important, because of its use in Bayes' formula. (Without it the sum of the probabilities of all possible hypotheses will not converge to one.) The invariance theorem for K(x) is similar to that of C(x). Now how can this measure be useful in extrapolating a sequence? First we will take a brief look at how Bayes' formula requires a prior probability.

6.4 Bayesian inference

We will start with a hypothesis space that consists of a countable set of hypotheses which are mutually exclusive, *i.e.*, only one can be right, and exhaustive, *i.e.*, at least one is right. Each hypothesis must have an associated prior probability $P(H_n)$ such that the sum of the probabilities of all hypotheses is one. If we want the probability of a hypothesis H_n given some known data D then we can compute that probability with Bayes' formula:

$$P(H_n \mid D) = P(D \mid H_n)\, P(H_n) / P(D)$$

where $P(D) = \sum_n P(D \mid H_n)P(H_n)$. This formula determines the *a posteriori* probability $P(H_n \mid D)$ of a hypothesis given the data, *i.e.*, the probability of H_n modified from the prior probability $P(H_n)$ after seeing the data D. The conditional probability $P(D \mid H_n)$ of seeing D when H_n is true is forced by H_n, *i.e.*, $P(D \mid H_n) = 1$ if H_n can generate D, and $P(D \mid H_n) = 0$ if H_n is inconsistent with D. So when we consider only hypotheses that are consistent with the data the prior probability becomes crucial. Because for all H_n where $P(D \mid H_n) = 1$ the posterior probability of H_n will become:

$$P(H_n \mid D) = P(H_n) / P(D)$$

Now let us see what happens when we apply Bayes' formula to an example of Solomonoff's inductive inference. In this example we only consider a discrete sample space, *i.e.*, the set of all finite binary sequences $\{0,1\}^*$.

What we want to do is, given a finite prefix of a sequence, assign probabilities to possible continuation of that sequence. What we do is, given the known data, make a probability distribution of all hypotheses that are consistent with the data. So if we have a sequence x of bits, we want to know what is the probability that x is continued by y. So in Bayes' formula:

$$P(xy \mid x) = P(x \mid xy)P(xy) / P(x)$$

We can take $P(x \mid xy) = 1$ no matter what we take for y, so we can say that:

$$P(xy \mid x) = P(xy) / P(x)$$

Hence if we want to determine the probability that sequence x is continued by y we only need the prior probability distribution for P(xy). Solomonoff's approach is ingenious because he first identifies x with the computer programs that can generate a continuation of x by a string y. In this way the a priori most probable continuation y can be determined in two ways: y is the next element that is predicted, *i.e.*, generated, by the smallest Turing machine program that can generate x; or the string y is predicted that is generated by most of the programs that can generate x.

So we can define the prior probability of a hypothesis in two different ways. We can give the shortest program the highest prior probability and define the probability of xy as:

$$P_{K(xy)} := 2^{-K(xy)}$$

i.e., the length of the shortest program that computes xy as the negative power of two (Li and Vitányi 1990, pp.216). Or we can define $P_{U(xy)}$ as the sum of $2^{-l(p)}$ for every program p (so not only the shortest) that generates xy on a reference universal prefix machine (Li and Vitányi 1993, pp.356). The latter is known as the Solomonoff-Levin distribution. Both have the quality that the sum of prior probabilities is equal to or less than one, *i.e.*,

$$\sum_x P_{K(x)} \leq 1 \text{ and } \sum_x P_{U(x)} < 1$$

However, it can be shown that if there are many ‘long’ programs that generate x and predict the same y, then a smaller program must exist that does the same. And it is proved that both prior probability measures coincide up to an independent fixed multiplicative constant (Li and Vitányi 1993, pp.357).

So we can take the Kolmogorov complexity of a sequence as the widest possible notion of shortness of description of that sequence. And if we interpret shortness of description, defined by Kolmogorov complexity, as a measure for parsimony, then the Solomonoff-Levin distribution presents a formal representation of the conceptual variant of Ockham’s razor, since the predictions of a simple, *i.e.*, short, description of a phenomenon are more probable than the predictions of a more complex, *i.e.*, longer, description.

6.5 Description length

While both the Kolmogorov and Solomonoff-Levin measure are not computable, there are computable approximations of them. It is demonstrated that several independently developed inductive methods actually yield computable approximations of Solomonoff’s method. I will first demonstrate this for Rissanen’s minimum description principle (MDL), *cf.* Li and Vitányi (1993).

Rissanen made an effort to develop an inductive method that could be used in practice. Inspired by the ideas of Solomonoff he eventually proposed the minimum description length principle. This principle states that the best theory given some data is the one which minimizes the sum of the length of the binary encoded theory plus the length of the data, encoded with the help of the theory. The space of possible theories does not have to consist of all possible Turing machine programs, but can just as well be restricted to polynomials, finite automata, Boolean formulas, or any other practical class of computable functions.

To derive Rissanen’s principle I first need to introduce a definition of the complexity of a string given some extra information, which is known as *conditional* Kolmogorov complexity:

Definition 2 *Conditional Kolmogorov complexity.* The conditional Kolmogorov complexity K_T of x, relative to some universal prefix Turing machine T(p, y) with program p and additional information y is defined by:

$$K_T(x \mid y) = \min\{ l(p) : p \in \{0,1\}^*, T(p, y) = x\}$$

Or $K_T(x \mid y) = \infty$ if such p does not exist.

This definition subsumes the definition of (unconditional) Kolmogorov complexity when we take y to be empty. Now, Rissanen’s principle can elegantly be derived from Solomonoff’s method. We start with Bayes’ theorem:

$$P(H \mid D) = P(D \mid H)\, P(H) / P(D)$$

The hypothesis H can be any computable description of some given data D. Our goal is to find an H that will maximize $P(H \mid D)$. Now first we take the negative logarithm of all probabilities in Bayes equation. The negative logarithm is taken because probabilities are smaller or equal to one and we want to ensure positive quantities. This results in:

$$-\log P(H \mid D) = -\log P(D \mid H) - \log P(H) + \log P(D)$$

When we consider $P(D)$ to be a constant then maximizing $P(H \mid D)$ is equivalent to *minimizing* its negative logarithm. Therefore we should minimize:

$$-\log P(D \mid H) - \log P(H)$$

This will result in the Minimum Description Length principle if we consider that the probability of H is approximated by the probability for the shortest program for H, *i.e.*,

$$P(H) = 2^{-K(H)}$$

Therefore the negative logarithm of the probability of H is exactly matched by the length of the shortest program for H, *i.e.*, the Kolmogorov complexity $K(H)$. The same goes for $P(D \mid H)$ and hence we should minimize:

$$K(D \mid H) + K(H)$$

This amounts to MDL principle, *i.e.*, minimizing the description or program length of the data, given the hypothesis, plus the description length of the hypothesis (Li and Vitányi 1990, pp.218). To make this principle practical all that remains is formulating a space of computable hypotheses that together have a prior probability smaller or equal to one, and searching this space effectively. It has been shown in several applications that this approach is an effective way of learning (Li & Vitányi 1993, p.371).

6.6 Cognitive models

Let us look at the simplicity bias of BACON.2 (BACON.2 is not representative for the other BACON programs. I discuss the simplicity bias of the other BACON's in (van den Bosch, 1994). BACON.2 will always construct the simplest consistent law in its range of search. The method it uses is called the differencing method. With this method BACON.2 is able to find polynomial and exponential (polynomial) laws that summarize given numeral data (Langley et al. 1987). One could define the simplicity bias of BACON.2 as follows:

Definition 3 *Simplicity bias in BACON.2* The simplicity of a polynomial decreases with the increase of the polynomial's highest power. A variable power is gathered to be a simpler polynomial than a polynomial with a high constant degree.

Langley et al. give no epistemical reason for preferring simplicity. However, after discussing simplicity in Thagard's PI I will argue that the simplicity bias of BACON.2, as defined above, is justified.

Thagard's account of the simplicity of a hypothesis does not depend on the simplicity of the hypothesis itself, but on the number of auxiliary hypotheses that the hypothesis needs to explain a given number of facts (Thagard, 1988). In PI, Thagard's cognitive model of scientific problem solving, discovery and evaluation are closely related. Simplicity plays an important role in both.

PI defines two hypotheses to be co-hypotheses if they were formed together for an explanation by abduction. From the number of co-hypotheses and the number of facts explained by a hypothesis its simplicity is calculated according to the following definition:

Definition 4 *Simplicity in PI.* The simplicity of hypothesis H, with the number of c co-hypotheses and with the number of f facts explained by H, is calculated in PI as

$$\text{simplicity(H)} = (f - c)/f \text{, or zero if } f \leq c.$$

To determine the best explanation PI considers both consilience (*i.e.*, explanatory success, or in PI; number of facts explained) and simplicity. This is no difficult decision if in one of the dimensions the value of one of the explanations is superior to that of the others while the values in the other dimension are more or less equal. If both explanations explain the same number of facts but one is simpler than the other, or if they are both equally simple, but one explains more facts than the other, then there is no difficult choice. But when *e.g.*, the first theory explains most facts while the second is the simplest, that conflict seems to make the choice more difficult. In that case PI computes a value for both hypotheses according to the following definition:

Definition 5 *Explanatory value in PI.* The explanatory value of hypothesis H for IBE is calculated in PI as

$$\text{value(H)} = \text{simplicity(H)} \times \text{consilience(H)}.$$

In this way PI can pick out explanations that do not explain as much as their competitors but have a higher simplicity or explain more important facts. It also renders *ad hoc* hypotheses useless because if we add an extra hypothesis for every explanation then the simplicity of that theory will decrease at the same rate as its consilience increases.

One feature of IBE in PI is that its valuation formula admits of a much simpler definition which easily follows from the definitions of simplicity and the value of a hypothesis as given above.

Theorem 1 For IBE the explanatory value of a hypothesis H, with the number of c co-hypotheses and f facts explained, can be calculated in PI as

$$\text{value(H)} = f - c \text{, or zero if } f \leq c.$$

Thagard does not satisfactorily argue why we should prefer this kind of simple hypotheses. In its defense he only demonstrates that several famous scientists used it. But he did not show that simplicity promotes the goals of inference to the best explanation, like truth, explanation and prediction.

6.7 Computable approximation

I will now compare Rissanen's minimum description length principle (MDL) with BACON.2, and with inference to the best explanation (IBE) as implemented by Thagard in PI. For BACON.2 Rissanen's principle suggests an improvement because in the case of noisy data, BACON.2 would probably come up with a polynomial as long as that data, while it could construct a much simpler one when it employs and encodes deviations from the polynomial as well.

An important difference between Rissanen's principle and BACON is that the former requires to search the whole problem space, while the latter searches it heuristically. But BACON's search is guided by the same patterns that eventually will be described by a law. However, a heuristic search, like BACON's, can be aided by Rissanen's principle. Actually BACON does search for a minimal description, but it does not try to minimize it, *i.e.*, if BACON finds a description, it halts, and will not search for a shorter one.

BACON.2 determines the shortest polynomial that can describe a given sequence. No Turing machine can be constructed that needs a shorter description for a more complex polynomial. It can be demonstrated that a polynomial formula with an exponential term has a shorter description than a polynomial formula without an exponential term that describes the same sequence. BACON.2's method always finds the simplest polynomial that exactly fits the data. So I will make the following claim:

Claim 1 The polynomial constructed by BACON.2 with the differencing method, based on a given sequence x is the polynomial with the shortest description that exactly describes x, if x can be described at all with a polynomial.

The validity of this claim can be derived from the differencing method. Every preference of BACON.2 between two polynomials that are compatible with the data is in agreement with the minimum description length principle. However, MDL can seriously improve BACON.2 by including a valuation of a description of the sequence, given a possible polynomial. A shorter description of the sequence may result when deviations from a possible polynomial are encoded as well.

In Thagard's explication of inference to the best explanation in PI, the simplicity of a hypothesis is determined by the number of additional assumptions or co-hypotheses that the hypothesis needs for its explanations. Rissanen's MDL accounts for the importance of auxiliary hypotheses as well. MDL requires that we minimize the sum of the description of an explaining hypotheses K(H) and the description of the data with the aid of the hypothesis K(D | H).

If an hypothesis can explain something right away the description of the data is minimal, while if the hypothesis requires additional assumptions, then the description of the data will be longer. So, Thagard's simplicity satisfies at least one of the requirements of MDL. Hence I want to make the following claim and argue for its plausibility.

Claim 1 In a case of equal consilience, the explanation that will be selected by IBE in PI will provide a shorter description of the facts, given the explanation, or at least no longer description with respect to the available alternatives.

This claim follows easily from the theorem stating that PI's IBE values hypotheses by subtracting the number of co-hypotheses c from the number of f explained facts, *i.e.*, f minus c. Two hypotheses that are of equal consilience explain the same number of facts, in which case the hypothesis with the least number of co-hypotheses is preferred. Hence, assuming that every such extra assumption is of about equal length, the simpler hypothesis will provide a shorter description of the facts given the hypothesis. However, if both have the same number of co-hypotheses, then PI can not make a choice, because both will provide a description of reasonably similar length.

6.8 Best hypothesis

One question may now come to mind: will the Solomonoff approach yield a unique preference when several simple hypotheses are compatible with the data? It seems possible that more than one theory or program, consistent with given data, can be of the same length. So in that case we cannot make a decision based on a simplicity consideration, because all alternatives are of equal simplicity.

To answer that criticism we first have to distinguish between the next symbol y that is predicted given a sequence x and the different programs p that can generate a prediction. Our goal can be a correct prediction y, given x, or a correct explanation of x. In the case that we want a correct prediction, if two programs are of the same length it may turn out that both predict something else. However, Solomonoff's method supplies two ways to solve this dilemma.

The first is the universal Solomonoff-Levin distribution with which probabilities can be assigned to different continuations of a sequence. A given prediction y not only receives a higher probability if it is predicted by a short program, but also if numerous programs make the same prediction. So if there is more than one shortest program, the prediction of the program that predicts the same as numerous other longer programs is preferred.

The second way out of the dilemma is in the situation when the given amount of data x is very long. It can be proved that in the limit all reasonable short programs will converge to the same prediction, so you can pick any of them. This feature of the Solomonoff approach is nice for practical and computable approximations. Because you can make reasonably good predictions with a given short program that may not be the shortest one.

However, if you value the best *explanation* of a given amount of observations, then you will not be satisfied by a grab bag of possible hypotheses that may not even make the same predictions. Scientists that want to understand the world usually look for *one* best explanation. In this situation a case could be made for the simplest hypothesis as the best explanation. But with such an aim the Solomonoff approach seems troublesome. Because you can never know whether a given short program that computes x is also the shortest program possible. Because the only effective way to do so is to test all possible Turing machines one by one to see if they generate x. But any of those possible Turing machines may never halt and there is no way to find out whether it ever will. You may put a limit to the time you allow the machine to run before you test the following one. But a shorter program can always be just one second further away.

The philosopher Ernst Mach once made the claim that the best thing that science could do is to make predictions about phenomena, without explaining the success of such predictions by the (ontological) assumptions of the possible hypotheses. However, the best explanation, and hence possibly the simplest program, can be seen as the ultimate goal of science. And a nice property of the present kind of simplicity is that we can measure our progress. We may not have an effective, *i.e.*, computable, method to establish whether a hypothesis is *the* simplest but given a large amount of data we *can* establish the relative simplicity of any two hypotheses that yield the data.

6.9 Conclusion

I will try to state my conclusion in one sentence, but nevertheless it is probably not the shortest description of that conclusion: in scientific discovery it is rational to prefer those hypotheses, that, given discovered alternative hypotheses, amount to the shortest computational description of known data, because they provide more probable predictions. This approach to learning and discovery generalizes the rational pretension of the logical and cognitive models of discovery that prefer minimal or simple explanations. So to answer the specific questions of this thesis, we have:

Question 1 What is the structure of a scientific theory? In the computational approach a theory consists of a universal Turing machine, together with a program for that machine. The data that is explained by the theory is the result of a description of that data that can be generated by a particular program that can make predictions. So, a string P describing data and predictions is the output of a computation of a computer TM and program H, *i.e.* TM(H) = P. Both the logical and the cognitive models can be subsumed within this general scheme.

Question 2 What is the process of scientific reasoning? The process of reasoning in machine learning is summarized in Table 6.1. A string P describes given data of a phenomenon p, given a way of representation. The task of explanation is to find a short program H* that can generate that string, given a computer T. This task is uncomputable, in the sense that there is no algorithm that can guarantee to find that program. Yet it can be approximated heuristically. The shortest program has the highest a priori probability. Given a Turing machine TM, a program H and earlier data P a prediction P* and posterior probability can be computed.

Problem	Start	Background	Process	Goal	Goal properties
Explanation	String P	TM	Approximation	H*	TM(H*) = P Prior probability
Prediction	Program H	TM, String P	Computation	P*	TM(H, P) = P* Posterior probability

Table 6.1: Short overview of inferences discussed in this chapter

Question 3 What is the route between theory and experiment? The theoretical route between theory and experiment can in this approach also be summarized in six steps. (For comparison, I added the logical version of this process between brackets)

1. Observe phenomenon p: $p_m, \ldots, p_n$
2. Describe p: $(P_m \rightarrow P_n)$
   $P_{m,\,n}$: = string
3. Explain p: $(H^* \models P_m \rightarrow P_n)$
   $TM(H^*) = P_{m,\,n}$
4. Predict p: $(H \models P_i^* \rightarrow P_j^*)$
   $TM(H, P_i^*) = P_j^*$
5. Intervene p: do p_i^*
6. Observe p: see p_j^*?

In step 1. a phenomenon is observed. This phenomenon is described by a string that represents the observation, given a manner of representation. An hypothesis H* is searched in step 3. in the form of a short program for a Turing machine, such that the program can generate the string and possible continuations of that string. Based on the given manner of representation, the program for H* and the string representing the observed data, the Turing machine can make a prediction by generating a continuation of the string, in step 4. Based on prediction an intervention and observation close the circle. The observation of new data does not necessitate a new hypothesis as long as the description of the new data plus the old hypothesis is still the shortest available description. New data do change the a posteriori probability of predictions.

In the next part of this thesis I will analyze a scientific practice to find out how that practice compares with the epistemological theories addressed in this part.

* * * *

Part III Neuropharmacology

What is the rational use of theory and experiment in the process of scientific discovery, in practice? In this part I discuss a case study and model of the rational use of theory and experiment in the practice of drug research for Parkinson's disease, as introduced in Chapter 3, in more detail. First I survey how the effects of drugs for Parkinson's disease are explained by the dopamine **theory** (Chapter 7). Then I report on the use of theory and experiment in **practice** (Chapter 8). I finish this thesis by discussing a model of both the dopamine theory and the studied practice of **discovery** (Chapter 9).

Chapter 7

Theory

7.1 Introduction

A short description of a theory and a practice in neuropharmacology, was introduced in Chapter 3 of this thesis. This third part provides a more detailed description and analysis of that same theory and practice of discovery.

The specific question for this part is: How are theory and experiments used in the practice of drug research for Parkinson's disease? To answer this question I will first survey the literature on the dopamine theory of Parkinson's disease in more detail. The particular question of this chapter is: how are Parkinson's disease and the effect of known drugs explained by theory?

Parkinson's disease is believed to be mainly caused by a deficiency of dopamine. Dopamine is a neurotransmitter, a chemical messenger between nerve cells in the mammalian brain. In this chapter I explore how dopamine is exactly related to Parkinson's disease, and how theory about that relation is used to understand the function of drug interventions for Parkinson's disease. Before discussing pharmaceutical interventions I will first discuss the dopaminergic cell and the basal ganglia in some detail to understand the rationale of these treatments.

In section 7.2 I start with a general introduction to Parkinson's disease. I go into the basics of the dopaminergic nerve cell in section 7.3. Then, in section 7.4, I go into the basal ganglia, the neural circuitry that partly controls voluntary movement, and how a defect in it causes parkinsonian symptoms. I end this survey of Parkinson's disease literature with a short overview of a selection of therapeutic drug interventions in section 7.5.

7.2 Parkinson's disease

People with Parkinson's Disease suffer from a motor behavior impairment, usually at an older age. The primary symptoms include: muscular rigidity, resting tremor, difficulty with movement initiation (bradykinesia), slowness of voluntary movement, difficulty with balance, and difficulty with walking. This disease was named after the English MD. James Parkinson, who in 1817 was the first person to describe these symptoms as 'the shaking palsy'. (Bernstein, 1995; Wichmann & DeLong, 1993)

Dopamine deficiency

More than a century later, one believes that the cause of the disease is a dopamine deficiency in the basal ganglia of the brain. Dopamine (DA) is a neurotransmitter, a chemical messenger in the nervous system, see Figure 7.1. In Parkinson's disease the neural cells which produce dopamine, the dopaminergic cells, deteriorate. When these neurons start to disappear, the normal rate of dopamine production decreases. It was discovered that when the degeneration of dopaminergic cells is more than 70-80%, Parkinson's symptoms start to appear. Next to Parkinson's disease's primary symptoms mentioned above, a patient may also start to suffer from secondary symptoms which include: depression, senility, postural deformity, and difficulty in speaking.

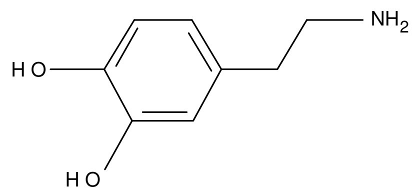

Figure 7.1: Structure diagram of dopamine

Diagnosis with L-dopa

It is difficult to diagnose Parkinson's disease in an early stage. The earliest symptoms may be non-specific, such as weakness, tiredness, and fatigue. So the disease may be unrecognized for some time. Today there are no conclusive tests for Parkinson's disease, yet there are several methods for evaluating its possible presence.

A first diagnosis is based on an evaluation of the presence and severity of the primary symptoms. If this test is significant, a trial test of anti parkinsonian drugs may be used to further diagnose the presence of Parkinson's disease. This test is usually performed with L-dopa. L-dopa is a precursor in the biosynthesis of dopamine in nerve cells, and causes the remaining dopaminergic cells to increase the production of dopamine. If the patient fails to benefit from L-dopa, the diagnosis of Parkinson's disease is questionable.

Parkinsonism

Computed tomography (CT) or magnetic resonance imaging (MRI) scans of the brain may be helpful in ruling out other diseases whose symptoms resemble Parkinson's disease. These diseases may include other neurological disorders leading to parkinsonian symptoms. Such symptoms can be caused by a brain tumor, repeated head trauma, or prolonged use of certain drugs. Such a condition is referred to as Parkinson's syndrome, or atypical Parkinson's. These kinds of parkinsonisms should not be confused with Parkinson's disease proper.

MPTP model

The cause of Parkinson's disease is still unknown. There is one known viral infection that damages the extra pyramidal nervous system and causes Parkinson's disease indirectly. However, the majority of sufferers were young people with different symptoms than we usually see in Parkinson's disease. Most of these cases resulted from an epidemic in the 1920's. More recently it was discovered that several young people who developed parkinsonian symptoms had used an illegal synthetic drug that was contaminated with the compound MPTP. It was found out that this compound is metabolized in the brain to a toxin that damages the extra pyramidal nervous system, causing a rapid decay of dopaminergic cells. Consequently it was hypothesized that Parkinson's disease is caused by an environmental toxic agent like MPTP. Yet, no toxin that has this effect other than MPTP is found in Parkinson patients. MPTP is now used in animal studies to understand how it causes these symptoms, which might lead to a better understanding of Parkinson's disease.

7.3 Dopaminergic cells

Research on Parkinson's disease focuses on the function of dopamine. This neurotransmitter is synthesized in the presynaptic terminal of a dopaminergic nerve cell by several metabolic pathways (see Figure 7.2 and Cooper, Bloom & Roth, 1996, pp. 293-351). First tyrosine in the cell is converted to L-dopa with the help of the enzyme tyrosine hydroxylase (TH). L-dopa in turn is converted into dopamine by the enzyme aromatic amino acid decarboxylase (AADC). The synthesized dopamine molecules in the presynaptic terminal are then taken up by synaptic vesicles. After the dopamine is released from the vesicles into the synaptic cleft, the remaining molecules are taken back into the synaptic terminal by transporters in the membrane. There they are transported back into vesicles or broken down to DOPAC by the enzyme monoamine amine oxidase type B (MAO-B) (Vermeulen, 1994).

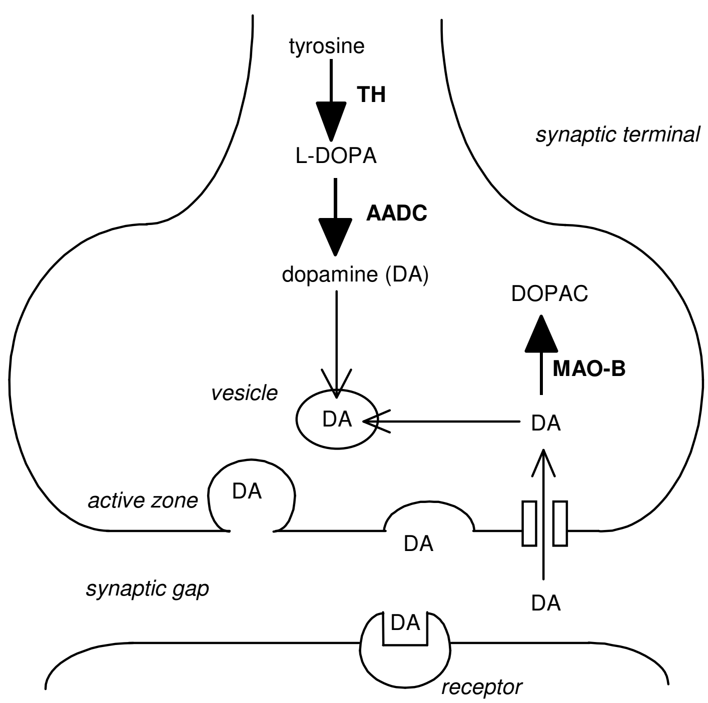

Figure 7.2: Prototypic dopaminergic terminal with cycle of synthesis, storage, release and removal of dopamine.

The signal to open or close ion-pumps is not determined by the chemical properties of a transmitter alone. The same transmitter chemical, *e.g.* dopamine, can both inhibit and excite other neurons, depending on the properties of the receptor it stimulates. Stimulated neurotransmitter receptors influence the membrane potential of a neuron directly or indirectly by various different mechanisms. There are ion channels with special receptor areas that directly bind with a transmitter. When bound to a transmitter these channels undergo a change that opens the channel immediately. The second type of receptors gate ion channels indirectly with a second messenger system. A transmitter bound to such a receptor causes in several steps the release of regulatory proteins within the cell membrane, that act on a family of ion channels.

7.4 Basal ganglia

Post mortem examinations of patients with Parkinson's disease revealed that parts of their brain were pathologically changed. This led to the believe that this part, called the basal ganglia, plays an important role in controlling voluntary movement. It was shown that signals from the cortex are led through the basal ganglia, to the thalamus, which influences motor control centers in the brain. (Côté & Crutcher, 1991)

Extrapyramidal system

The basal ganglia became known as a component of the so-called extrapyramidal motor system, which was first presumed to operate independently of the pyramidal or corticospinal system. However, today it is known that both systems are interconnected, and cooperate. Furthermore, other parts of the brain are shown to play a part in voluntary behavior as well, and the basal ganglia also have a role in cognitive functioning.

The basal ganglia themselves are a conglomeration of five distinguishable interconnected nuclei. They are called the:

- globus pallidus, internal (GPi) and external segment (GPe)
- subthalamic nucleus (STN)
- substantia nigra, pars compacta (SNC) and reticalata (SNR)
- striatum, consisting of caudate nucleus and putamen

From the cortex there is a direct and an indirect signal pathway through this conglomeration, maintained by circuits that use different neurotransmitters, such as GABA, glutamate, enkaphalin and substance P. There is a delicate balance between these two pathways that is partly maintained by dopamine release from the substantia nigra to the striatum. Dopamine release inhibits the indirect pathway by stimulating dopamine D2-receptors, and excites the direct pathway by stimulating the dopamine D1-receptor (see Figure 7.3A, Timmerman 1991, Vermeulen, 1994). The thickness of the arrows represents the strength of the signal. In the case of Parkinson's disease the indirected path is less inhibited, so becomes stronger. The direct path will lack amplification and will become weaker.

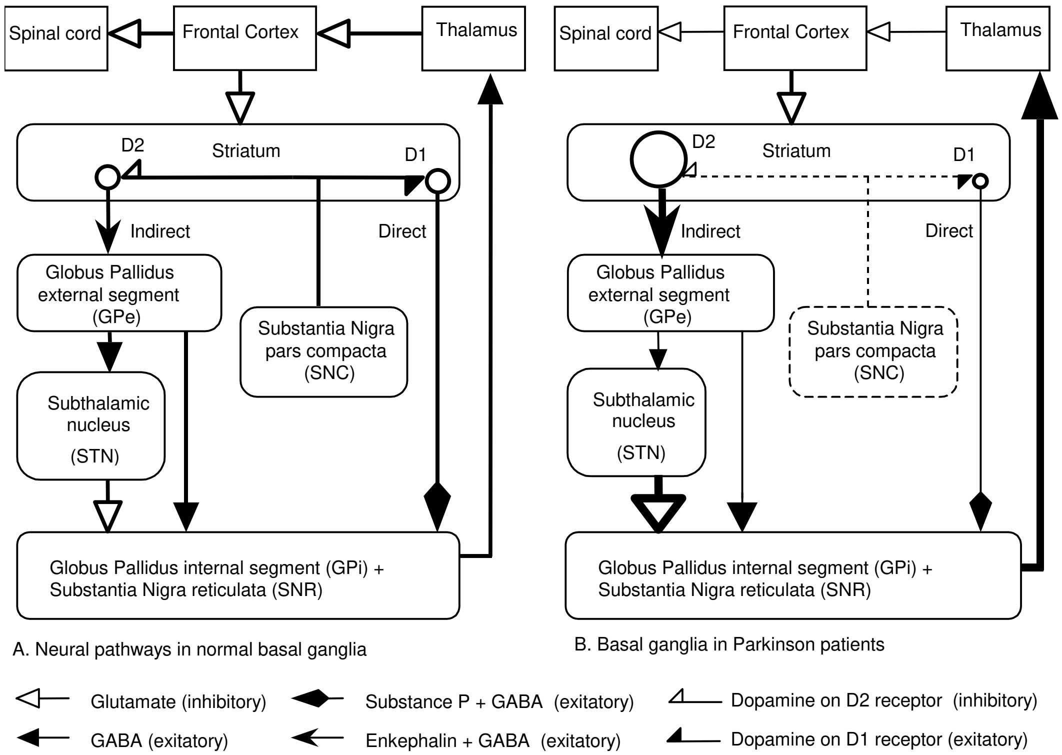

Figure 7.3: Major neural pathways in normal and Parkinsonian basal ganglia

Substantia nigra

In postmortem studies it was discovered that the substantia nigra (meaning "black substance"), had lost its pigment in Parkinson patients. Subsequent studies showed that dopamine levels in the striatum were drastically reduced. Because the basal ganglia contains most of the dopaminergic neurons of the brain, these observations suggested that the dopaminergic pathway between the striatum and substantia nigra is degenerated in Parkinson's disease patients. It was theorized that the depletion of dopamine disbalances the direct and indirect pathways from the striatum, which causes the thalamus to be overstimulated. As a result the frontal cortex is less activated, which would contribute to the Parkinsonian symptoms (see Figure 3B).

7.5 Drug treatments

L-dopa

Given the observations in the basal ganglia in the early1960's Birkmayer and Hornykiewics reasoned that it would possibly help Parkinson patients if the level of dopamine was restored to normal levels. It is not possible to administer dopamine itself as a drug because it will not pass the blood-brain barrier between the blood vessels and neurons. However, L-dopa, the precursor in the synthesis of dopamine will. So they reasoned they could boost the dopamine production up to higher levels by providing the few remaining healthy dopaminergic neurons with large amounts of extra L-dopa. (Côté & Crutcher, 1991; Vermeulen, 1994)

The first tests led to a successful initial remission of the symptoms. Yet this positive effect was countered by serious side effects such as nausea, vomiting, blood pressure changes, and collapse. This could be explained by the fact that the enzyme AADC, which converts L-dopa to dopamine, is also present in the liver, kidney and many other places in the body. So while the dopamine levels in the striatum became more normal, the extra dopamine production disturbed chemical balances elsewhere in the body.

AADC inhibition

After further studies it was demonstrated that the effect of the L-dopa treatment was enhanced when the dose of L-dopa is increased more gradually. So the focus of research became the reduction of the side effects. In the early 1970's the first AADC inhibitors that could not pass the blood-brain barrier were introduced. This made it possible to increase dopamine levels in the brain only, because the conversion of the extra L-dopa in the peripheral organs could be inhibited selectively.

MAO-B inhibition

Another way to increase dopamine levels is to block the enzyme MAO-B that is converting dopamine to DOPAC. It is demonstrated by studies that the administration of MAO-B inhibitors slows down the progression of Parkinson's disease, and increases the life expectancy.

It is argued that this slow down can also be explained by the hypothesis that Parkinson's disease is caused by a toxin similar to MPTP. It was shown that MPTP needs to be converted to MPP+ by the enzyme MAO-B to have its destructive effect. So if some toxin like MPTP causes the cell death in the basal ganglia of Parkinson's disease patients, the inhibition of MAO-B would slow down this process.

Yet it is also argued that the positive effect of MAO-B inhibition can be (solely) attributed to the effect that it inhibits the break down of dopamine, and hence increases the dopamine level.

L-dopa treatment only symptomatic

While L-dopa is the best available remedy to ease the lives of Parkinson patients, it is not even near a cure. Treatment that aims to increase dopamine levels turns out not to stop the further deterioration of dopaminergic cells, and hence does not work well in the long term. Long-term use of L-dopa frequently results in fading of the therapeutic effect and the development of serious side-effects, such as further motor impairment and psychiatric complications. Furthermore, while the lack of dopamine causes most of the Parkinson symptoms, Parkinson's disease patients also suffer a loss of noradrenergic and serotonergic neurons, which contributes to the disease as well.

Dopamine receptor agonists

To bypass the problem of the side effects of L-dopa treatment, research was initiated to synthesize compounds that would directly act on the dopamine receptors. These compounds, called receptor agonists, would take over the role of dopamine, so no administration of L-dopa would be needed. And hence the side effects induced by large amounts of L-dopa would be countered.

To date this ideal has not yet been reached. While long-term treatment with the available dopamine receptor agonists results in less dyskinesias, the therapeutic effect is less than that of L-dopa. And increasing the dose only leads to other serious side effects such as psychotic reactions. Better effects result from a combination of a low doses of L-dopa with an agonist.

There are also others reason for research into dopamine receptor agonists. It has also been put forward that long term treatment with L-dopa accelerates the degeneration of dopaminergic cells. This could be caused by the enhanced generation of toxic free OH-radicals through dopamine auto-oxidation (Vermeulen, 1994). The higher the amount of dopamine in the cell through extra L-dopa, or MAO-B inhibition, the higher the risk of toxication. If this claim is true, it is preferable to use receptor agonists.

Furthermore, synthetic agonists have the advantage that they can be made highly selective for a particular receptor. There are now five known types of dopamine receptors, and further knowledge of how they are integrated in neural circuits that regulate motor behavior may result in an agonist with less (but also different) side effects.

7.6 Conclusion

In this chapter I asked: How are theory and experiments used in drug research for Parkinson's disease, according to the literature? Theories about the neurophysiology and biochemistry of the brain are used to explain the pathology of Parkinson's disease, and the function of known drug interventions. In neuropharmacology theory serves to guide the search for new and better drugs. In this chapter I surveyed the dopamine theory of Parkinson's disease, and how theories about dopamine's metabolism and function imply suggestions for treatment. In the next chapter I survey part of a practice of research on Parkinson's disease.

* * *

Chapter 8

Practice

8.1 Introduction

How are theory and experiments used in the practice of drug research on Parkinson's disease? Several techniques are being used to search for new drugs and explore the activity of the basal ganglia. In this chapter I report on how new drugs are investigated and how experiments are being used to explore and test new drugs and the mechanisms of the brain at the Pharmacy Department of the Groningen University.

For my case-study I interviewed researchers Dr. B. Westerink and Dr. W. Timmerman. In the following sections I will report on their views and experimental work. Section 8.2 presents an overview of my interview with Dr. B. Westerink. Sections 8.3 to 8.5 report my more extensive interviews with Dr. W. Timmerman which I partly conducted while witnessing her work in the laboratory.

The numbered paragraphs in these sections are translations of a selection of the verbatim responses to my questions, which were reviewed and approved by the interviewees. They aim to present an objective picture of the researchers' views on their work. Off course, all errors of translation and interpretation remain my responsibility. The next chapter of this thesis will present a detailed analysis of the practice that is portrayed in this chapter. The paragraphs are numbered for ease of reference.

8.2 Investigating new drugs

Dr. B. Westerink is a senior researcher, conducting his work at the Pharmacy Department of the Groningen University. The following text reports his views in response to my questions about drug discovery in the context of drug research for Parkinson in general, and more specifically at his department. The interview was conducted in December 1996.

1 Drug experiments can serve to investigate how and why a drug has a particular effect, whereas that effect is often discovered by accident. In 1960 the mechanism of neurotransmission became better understood. In 1965 it was discovered that already known kinds of compounds had a neurotransmitter function. Carlson discovered that in Parkinson's disease dopamine was deficient.

2 By an accidental observation it was discovered that chlorpromazine, while it was being administered for a different reason, improves schizophrenia. By focused experiments on rats it turned out that that this drug had an effect on the amount of dopamine. The hypothesis was proposed that chlorpromazine blocks the dopamine-receptor, which would cause the brain to compensate by increasing the synthesis of dopamine. This hypothesis is accepted today.

3 Often you see that an effect of a drug is discovered by accident, by a secondary observation. This will then initiate further research to understand the specific function of a drug. Later it was discovered that chlorpromazine causes parkinsonism as a side effect. This suggested a relation between DA and Parkinson's disease. This hypothesis [as discussed in the former chapter] is a result of further experimental investigations. This hypothesis pointed to rational strategies for therapy and novel drug design.

4 One direction that is explored in Groningen is the development of selective DA-agonists. These are chemical variants of the structure of dopamine. Those variants are experimentally tested in vivo (on live animals) and in vitro (on samples of tissue in a test tube) for their biological activity on a receptor. As a reaction to an agonist a receptor can make Cyclic-AMP. The concentration can be measured and compared with the concentration that is released after contact with dopamine.

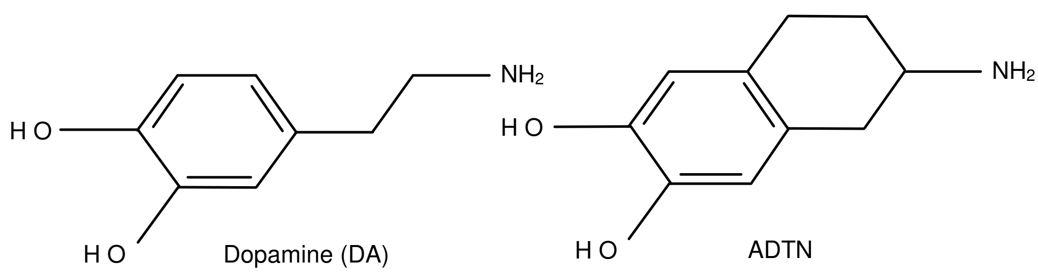

Figure 8.1: The chemical structure of dopamine and its variant ADTN

5 In 1977 a variant of DA was conceived by Prof. Horn (former professor at the Groningen University), called ADTN (Figure 8.1). This structure was the basis for further variants that were experimentally tested on four criteria:

1. The activity on the DA-receptor
2. Lipophilicity, the ability to cross the blood/brain barrier
3. Metabolism, its decomposition by enzymes
4. Selectivity, its affinity for D_1 and D_2 receptors

6 Suggestions for variants are based on experience and *fingerspitzengefühl.* It is hard to exactly predict what a receptor and enzymes will do with a compound. Yet the design of an antagonist is somewhat less difficult because it only has to obstruct a receptor, while an agonist has to fit and activate the receptor, like a key.

7 The NH_2-group of the best variant of ADTN was extended with two propyl-groups. This increased its lipophilicity so dramatically that it could even be administered by a band-aid on the skin. Removing a hydroxyl-group decreased its metabolism.

8 The experimental search also evolves the other way around. When a new receptor is discovered, its genetic expression can be used to clone it. These clones are then used by pharmaceutical companies to test all their created compounds for activity on that receptor. If one of the often more than 100.000 compounds is found to be active it can be the basis for a new drug lead.

9 Another strategy is using techniques from combinatorial chemistry to create thousands of variants of a compound at the same time and test them by rapid screening techniques. If activity is observed the compound that caused it is retrieved and will be studied to discover its structure.

10 A computational approach builds 3D models of receptors. These are used to aid the understanding of drug docking mechanisms [how a drug interacts with a receptor] by simulating and visualizing that process. Such simulations make predictions possible about how a protein folds and deforms.

11 If a new drug passes the criteria of the lab it will then be extensively tested. This is a process in three phases. The first phase tests for toxicity. It is administered to animals and later to volunteers. In the second phase the drug is given to volunteering patients to test its therapeutic strength. When it passes this barrier it goes into double blind testing and will be used in hospital trials. This is an expensive process, and yet the drug can still make victims, even when it passes all three stages. Genetically heterogeneous human beings are not the same as homogeneous mice. It is always possible that a slight genetic mutation will make a compound highly toxic for a particular group. Sometimes serious side-effects occur within isolated groups, *e.g.* in Finland or in Jewish families.

12 Newly created or discovered drugs are also used to explore biochemical mechanisms in the brain, both in normal as well as pathological conditions. This is another area of neuropharmacology. For Parkinson's disease the basal ganglia are of great interest.

8.3 Exploring the basal ganglia

In Groningen the nuclei called the basal ganglia were being studied by Dr. W. Timmerman and her students. The following text reports her responses to my questions about Parkinson's disease in general, and her experimental work on the basal ganglia in particular. I conducted these interviews in January and February 1997, and in September 1998.

In this section Dr. Timmerman talks about how the basal ganglia are involved in Parkinson's disease, how they are explored experimentally, and how knowledge about them can be used to treat Parkinson's disease. In Section 8.4 Dr. Timmerman talks about a specific experiment that was conducted during the interview. Section 8.5 reports her thoughts on interpreting data from experiments in general, and the conclusion of her experiments on the role of dopamine in the basal ganglia in particular.

13 The basal ganglia present the nuclei in the brain where the neural activity is abnormal in Parkinson's disease. When activity changes in the basal ganglia, all kinds of adjustments take place. Via the substantia nigra information is processed to other structures, to the thalamus, and then back to the pre-motor cortex. But we

do not know how, precisely. We also do not know exactly how information is processed from the basal ganglia to the periphery.

14 We know that the striatum processes information via a direct and indirect pathway to the SNR. From the nigra connections go further to the brain stem, and from there to the spinal cord. This can constitute a direct control of certain muscles. But there are also pathways going back via the thalamus to the cortex. So it is also possible that for example a change in activity of the corticospinal pathway is necessary for the deviation in behavior and motor control. I think it is a combined action. It is not just the basal ganglia and neocortex. The thalamus is involved, just like the cerebellum, which in turn also projects to the thalamus and the spinal cord. If the activity changes in the thalamus and the neocortex by input from the basal ganglia, then these changes can spread through the brain, making adjustments elsewhere.

15 Much is known about the anatomy of the basal ganglia. It is much more refined than is depicted in the model [see, Figures 3.1 and 7.3]. For example, it now seems that there are also dopaminergic neurons projecting to the Globus Pallidus (GP) [see Figure 8.2]. It also seems that the direct pathway has branches to the GP [see dotted lines in Figure 7.1].

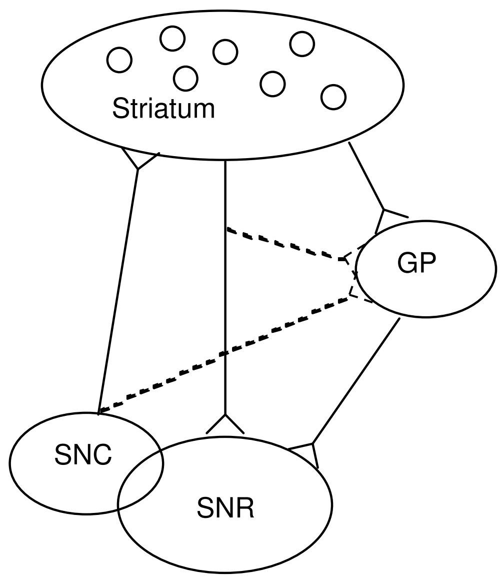

Figure 8.2: Schematic illustration of the basal ganglia by W. Timmerman

16 These pathways are discovered by means of tracing methods, *e.g.* by color coding and mRNA detection. Even so, it was discovered that the striatum is not a homogeneous structure. It is now known that it contains limbic patches in a matrix [see circles in Figure 8.2]. These areas specifically receive information from limbic areas in the cortex. From those patches specific information is processed onto the dopaminergic cells in the SNC. This is just the anatomy. These patches are chemically different, but not electrophysiologically.

17 In my study of the basal ganglia I specifically look at postsynaptic processes. I ignore anything that happens in the dopaminergic cell. I apply dopamine-agonists locally. In doing so I overrule the dopaminergic cell by activating post synaptic receptors, the receptors on *e.g.* the GABA-ergic and cholinergic cells. It is not my

problem how the endogenous dopamine is released by the cell, or how it originates from L-dopa or is broken down by MAO-B. Other people look at those specific processes. That is a research preference. Of course in the end it all has to fit together.

18 Anatomically you can look at one cell, one synapse, and you can identify projections and pathways. But for function you can learn from behavioral studies. These are often used as a measure for activity in the striatum. By local infusion of GABA-like and dopaminergic compounds in a certain part of the basal ganglia you can induce prototypical motor behaviors. By increasing and decreasing these compounds in different parts you can develop a concept of the role of GABA and DA on this level.

19 If you want to know what dopamine does in the brain then you can for instance give amphetamine, a compound that will induce the release of dopamine. If you administer it to a rat, it will show stereotypical behavior. It is a simple test that shows that dopamine is related to behavior.

20 But the question is: where is this dopaminergic effect mediated? Dopamine is not located in just one brain area. There is the nigro-striatal dopaminergic pathway, but there are also dopaminergic pathways that lead to the cortex and the accumbens. So you can specifically inject amphetamine in the cortex, striatum or accumbens. You will only see that specific stereotypic behavior if you inject amphetamine in the striatum. If you apply it in the accumbens you will mainly see locomotor activity, not stereotypical but an increase in locomotor behavior.

21 Behavior is a very accessible measure in experimental research. You have a cage, you have a rat, and you can start your research. So I start with that. In this way you can get new ideas about what a certain transmitter or pathway does in the brain, even though you are doing very little in the animal. You just look at simple behavior, is it running, is it stereotypical or not? You can also do very complicated experiments with learning models, then you explore different pathways.

22 In behavioral experiments you look at the end product of your injection. Another way is to measure the response directly in the brain in specific areas by inserting a microdialysis probe into the brain. In that way we can measure the direct effect of amphetamine, it releases an enormous amount of dopamine. The explanation for that effect is that the vesicles that contain dopamine fuse with the cell membrane, resulting in the release of dopamine in the synaptic cleft.

23 The research that I am doing aims to bridge the gap between our knowledge about the anatomy of the basal ganglia and pharmacology/physiology. From the anatomy we know that there is a connection between dopaminergic pathways and the striatum. But what is the effect of a dopamine transmitter on the striatal cells? That is still not entirely clear. It is an essential question.

24 In Parkinson's disease dopamine depletes in the striatum. That appears to be a major problem. To solve that problem you can administer L-dopa or dopaminergic agonists. But these turn out not to be ideal therapeutics because after a while side effects appear and they gradually loose their therapeutic effect. What you rather want to know is the function of dopamine in the striatum. If you have a better idea of its function it may be possible to use other, more specific compounds. So, it is essential to know what dopamine does in the striatum.

25 We have learned that there are two main subtypes of dopamine receptors in the basal ganglia, D_1 and D_2. So I work with compounds that are specific for those subtypes. But what is the function of those subtypes? In the literature this question has been asked many times. Should we just use a compound specific for D_2 or one that acts on both types, such as *e.g.* apomorphine? This compound has not been used for a while because of its many side effects. But it now seems to be a reasonable alternative.

26 Also L-dopa has D_1/D_2 affinity, simply because it results in more dopamine that acts on both subtypes. It is an ongoing discussion, what is ideal? You want to replace dopamine, you cannot use dopamine itself, so what do you need? Do you need to activate only one subtype so that you restore function, but not induce side effects? Is activation of D_2 receptors enough, or do you need a little bit of D_1 receptor activation too, and what is the ratio?

Searching a treatment

27 The problem of all dopamine-agonists is that they also have side effects in the periphery, in other places of the body like the heart and veins. You can counteract that by administering peripheral dopamine antagonists, like domperidome, together with a dopamine agonist. That will relieve side effects like nausea, but has its own side effects.

28 Another difficulty is that you have to find a proper dose that may differ per person. Too much DA-stimulation will lead to an over-activation that can induce *e.g.* spontaneous dystonias. [Dystonias are movement disorders in which sustained muscle contractions cause twisting and repetitive movements or abnormal postures.] Usually after about five years patients will be increasingly in an off-period. In an on-period a patient reacts positively to medicine. In an off period the reactions are either poor or hyper.

29 In theory you try to maintain a level of dopamine receptor stimulation by administering dopamine agonists. But when you apply a compound the sensitivity of the receptors changes. In Parkinson patients the dopamine receptors become hypersensitive as result of the dopamine depletion. By a process of up-regulation the number of receptors on cells increases. This changes, for instance, also the uptake of dopamine. There are all kinds of mechanisms that act as soon as something changes, to compensate for the change.

30 By administering an agonist you try to reestablish the situation that was normal before the degeneration of the dopaminergic cells. But you do not know what that situation was. In the clinical practice, different doses are tried until the patient's motor behavior returns to normal. But that dose might not be comparable to the amount of dopamine that was normally released before the degeneration. When you start medication the receptors are still hypersensitive. But that will change, and the induced effect will eventually decrease, so the dose should be adjusted.

31 There are methods to establish sensitivity. But there are also all kinds of compensation mechanisms on other levels than the dopamine system. Changes in dopamine induces changes in acetylcholine in the striatum, and also changes in GABA and glutamate. So, how to solve that problem, how to chart that system and how to restore it to normal?

32 The clinical studies are a kind of trial and error. The therapy is thought to be adequate when the patient responds well. Parkinson patients respond well to dopaminergic agonists, but also to cholinergic antagonists. There has been long discussion about an explanation for that effect. It is thought that there is a DA-acetylcholine balance in the striatum. When dopamine increases, acetylcholine decrease, and vice versa. That would explain why dopamine agonists and acetylcholine antagonists have a similar effect.

33 By doing basic experimental work it now appears that stimulation of D_2 receptors on cholinergic neurons does indeed inhibit the cell, explaining the balance. But, via D_1 receptors and via the cortex dopamine can also stimulate acetylcholine. So there is a delicate balance between an inhibitory and excitatory effect of DA on acetylcholine functioning. It can only be discovered by basic experimental research. How to incorporate such specific knowledge into the practice of treating Parkinson's disease is another problem.

34 Another approach is to study the effect of using NMDA antagonists. NMDA is a glutamate receptor subtype. Maybe we should use such a compound in combination with a dopaminergic therapeutic to create the optimal effect. However, the problem is that glutamatergic activation will influence the whole brain. You immediately interact with all kinds of other areas, so that will not be my best bet. Yet if you would understand how glutamate interacts with dopamine then you could judge this better. But given our current knowledge it is still a long way to go before we can easily infer what to do.

35 For my own research I want to know what the effects are of D_1 and D_2 receptor stimulation in a healthy situation. If you got a good idea of that, you can look at a lesioned model to verify if the effect is the same in the pathological situation. Is the interaction between glutamate and dopamine and dopamine receptor subtypes still traceable in the same way? If that is not the case then you must better establish what kind of compensation is involved after a dopamine cell lesion.

Using the model

36 At the moment the role of dopamine in the striatum is still a matter of debate. We have a model which claims that there are excitatory D_1 receptors on the direct pathway, being separated from the inhibitory D_2 receptors on the indirect pathway. But if you look at the literature, all kinds of gaps emerge in this story. It is pleasantly simple, but it completely lacks nuance.

37 For example, it is dubious whether D_1 receptors are located only on the direct path and D_2 receptors only on the indirect pathway. This claim alone is subject of enormous debate. There is a group of well known anatomists that claim that there is a division, based on studies of rats and monkeys with a dopaminergic cell-lesion. A way to discover the presence of receptors in pathways is by looking at messenger RNA. But with the same methods another group claims that D_1 and D_2 receptors are present on both pathways, with no absolute segregation.

38 My own data also do not fit the model. The model explains many findings but also leaves a lot of questions. But in the literature many authors appear to just treat the model as given, apparently without questioning it. This is something that intrigues me. Even though it does not fit the data well, it has gained enormous popularity. Why? I think that it is because the model fits the way you think that it

will work. It provides a prediction that you can easily understand. It is simple and it is beautiful to work with. That is why I think so many people just take it for granted without questioning it.

39 The reason that it is beautiful is the following. Upon activation of the cortex you get a glutamate activation in the striatum. Now if glutamate acts in a similar manner on both the direct and indirect pathway, you get a net reciprocal effect in the SNR, they counteract each other. The model shows that dopamine acts synergistically with glutamate stimulation via the D_1 receptor to increase the amount of GABA in the SNR, inducing an inhibition of the nigral activity. At the same time dopamine inhibits the indirect path that would increase nigral activity via the D_2 receptors, therefore diminishing the excitation of the nigral cells. So, dopamine let the activities of both pathways point in the same direction. It stimulates the direct pathway and inhibits the inhibition via the indirect pathway. The net result is a decrease of activation of the SNR. This is associated with behavioral activation. It increases the activation of the thalamus and brainstem, which coincides with all kinds of activity.

40 That is why it is beautiful, dopamine is a compound that facilitates activation. For example, with amphetamine you see stereotypical locomotion activity. You can understand that behavior using the model that says that an increase of dopamine results in SNR inhibition, which enables behavior activity.

8.4 Testing the model

In this section I report on how Dr. Timmerman used the basal ganglia model to conduct her own experiments in the laboratory. Part of the interview was conducted in the laboratory.

41 The model is subjected to heavy criticism. The first thing I did was to check whether a change in activity of striatal cells caused a change in the SNR. I infused glutamate agonists of several receptor subtypes and an immediate decrease of activity could be observed in the SNR. That means that apparently the direct route is stronger than the indirect route, as otherwise activation of the latter pathway would induce an increase in activity, given the model. So I tried several glutamate agonists to confirm the model.

42 After that we did a test with a D_1-agonist. The result was a very slight decrease. Although the effect was very limited, it would be in accordance with the model. However, application of a D_2-agonist induced a gradual but again very minor increase. If any, it does not fit the model. So is there a real segregation between the two pathways? The effects are hardly noticeable.

43 But is activation of the D_1 receptor always stimulating? In vitro studies never show a stimulation by D_1 receptor activation. If you prepare slices of the striatum and you apply a D_1 agonist you will not see a stimulation but an inhibition. That does not fit the model. So for me it is more like a model you work with, knowing that there is a lot more nuance to it. Also people that perform those in vitro studies never talk about this model. It does not fit their data, so why would they accept it.

44 What we know from other electrophysiological studies is that GABA-ergic neurons in the striatum are hardly active, under basal conditions. You can easily activate them with glutamate. We assume that dopamine modulates the glutamate-GABA interaction in the striatum. But if there is very limited activity in the striatum, a modulator will hardly be effective. So I thought, let's give a slight activation of the striatum by glutamate, and then let's see if we can make the modulating role of D_1 and D_2 agonists more apparent. The literature also implies that the role of dopamine depends on the influence of glutamate.

45 My presumption is, D_1 probably excitates, D_2 probably inhibits, possibly on different pathways. Can I confirm this, or cannot I? Well, I cannot confirm everything. Under basal conditions, without activation by glutamate, you can not speak of dopamine as a modulator, because there is nothing to modulate. That was my former study. Having finished that, I am now searching for a better start situation. That means I have to induce a slight glutamatergic activation locally in the striatum. I tried that, but it was difficult. You cannot have a nice constant activation because all kinds of other systems immediately try to compensate the increase in activity.

46 What I tried together with a student of mine, is to stimulate at the level of the cortex with a glutamate agonist, and look if this activation is noticeable in the SNR. You expect that the activation of the cortex will release glutamate in the striatum, that will consequently result in activation of GABA-ergic neurons. Depending on what pathway is the strongest, this should decrease or increase the activation of the SNR. So first we want to know which pathway dominates upon activation, but only to search a situation to again test the role of dopamine in the striatum.

47 After performing these studies it seemed that the cortex is not the best place to start the activation. So now we try to start with activating the thalamus. The thalamus projects both via the cortex and directly on to the striatum. That would create a general activation in the striatum. We have seen that if you infuse a glutamate agonist in the thalamus, just for ten minutes, then you will see a slight reaction. Yet we could not confirm this in later studies.

48 The suggestions to change experimental conditions are based on both the model and our former experiences. According to the model glutamate with D_1 receptor activation will increase the activation of the SNR, they amplify each other.

49 The test we are running now [Feb. 26, 1997] is to first infuse a D_1 agonist into the striatum. Secondly we will give a glutamatergic stimulation of the striatum to find out if D_1 cooperates with glutamate to induce an increase of GABA and hence an inhibition of the SNR. We want to find out if the presence of a D_1 agonist makes a difference. We have done this D_1 agonist infusion three times already. We have seen some reaction, but very little.

Performing microdialysis

50 In the laboratory we use brain dialysis probes. Such probes consists of a small glass tube with at the bottom a semi permeable membrane and at the top two extensions, an inlet and an outlet. If a fluid is infused via the inlet, diffusion into the surrounding tissue at the tip of the probe occurs. You can infuse compounds in this way, but you can also sample from this area. Depending on where the area of

interest is located in the brain, and what the dimensions are of this brain area you can make longer or shorter dialysis probes.

51 You can implant the probe in the brain in such a way that the tip is at a specific location. For our experiments we use Wistar rats. You can find a location in its brain using the atlas of the rat brain. Ours is falling apart because of its extensive use. The atlas portrays the whole brain from back to front in slices. We want to put our probe in the striatum. This area is relatively large, and both rats and human beings have two striata. It is a stretched out area that runs through a large part of the brain. To put in a probe you look at the coordinates of the map. These are standardized for a Wistar rat of 300 gram, and you look for particular blood vessels. The bregma at the center on top of the skull is a reference point. All brain slices portrayed in the map have a known distance from the bregma. For the striatum you look at the map that is just behind the bregma, the probe should be located 3.5 mm to the side, and 7 mm deep.

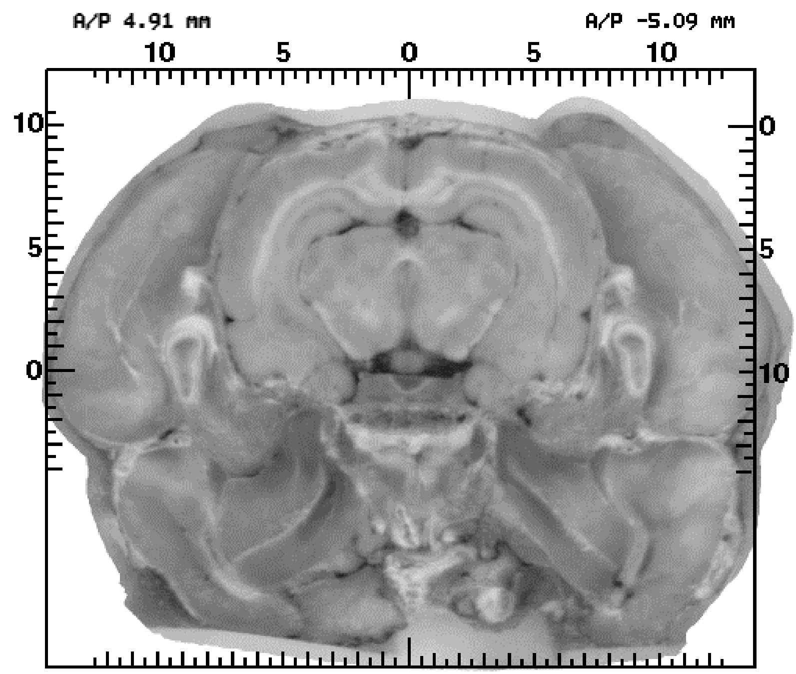

Figure 8.3: Example slice from the rat brain atlas

52 We place an anesthetized rat in a stereotactic apparatus, that clamps its scull by the ears and at the nose. Using this apparatus you can exactly determine a given location, using the bregma as a reference. When we find the given location and drill a small hole in the scull of the rat. After that we slowly lower a probe inside. We seal the probe with a screw and some cement. You prepare a rat a day in advance. When we add an electrode, to measure electrical activity, we always do this just before the actual experiment starts.

53 The inlet of the dialysis probe is connected to an infusion pump. Very slowly an ionic fluid is infused through the brain area, and it leaves the brain via the outlet tube. The fluid that comes out reflects the compounds that are present in that location at that moment. That fluid is guided through a system that analyzes a sample every given time-interval. That system can be set up to measure particular compounds, such as amino acids, dopamine, noradrenaline, etc. The compounds are separated in a column from where they are guided to a detector. For example, for amino acids we use a fluorimeter that registers the degree of fluorescence that is detected and plots those values against time.

54 In the case of our current experiment we have put in two microdialyis probes, one in the thalamus and one in the striatum. Additionally, in the SNR we place an electrode, which is an isolated wire with a small uncovered tip. This tip can measure electric activity outside the cell, it is still to large to measure intracellularly. In this way we can make an extracellular recording of action potentials. You pick up those signals from one or two nearby cells. You can determine what kind of cell you are measuring by looking at characteristics of the action potential. When we lower the electrode we actually try to find a particular cell type by looking at the kinds of signals, given descriptions in the literature. The SNR neurons are described as being tonically active, displaying a high firing frequency, and exhibiting a nice thin action potential. We have to find the correct type since other cell types may also be present in the same area. In this experiment it is easy because the SNR mostly contains the same type of cells.

55 To diagnose the cell signal type we use a computer program that records templates of signals we are interested in. You record an example of a signal and tell the program to start looking out for those types. It distills the signals from the noise. We know that an SNR action-potential lasts about 0.7 milliseconds. Any signal that takes longer will be ignored. Depending on the location of a cell and its connections, it displays a particular electrical activity. The activity depends on incoming signals from other cells or it can fire spontaneously. In our experiment we know that the cell fires about 20 to 40 times a second. An extra condition for this experiment is that the firing frequency remains stable in time. If the activity we monitor deviates from those conditions we start looking for another cell by moving the electrode again. To find good cell activity one has to learn; it will take time, patience and frustration.

56 When all conditions are met we start the experiment. We need a good base line, a good firing frequency, the activity needs to be tonic, the rat should be well anesthetized. All conditions should remain stable for half an hour. Then we record a base line of the activation for ten minutes, and start the fluid infusion. That is all still part of the preparation. When all goes well up to that point we decide to start the actual experiment or wait or look for another cell. If everything looks good we do not touch the rat anymore and start the experiment. The only thing that remains is to change the syringe from the one containing an isotonic fluid to the one that contains the isotonic fluid with the drug to be applied and start the drug infusion. The compound will enter the brain and now all we need to do is see what will happen.

57 We have done a test with a D_1 agonist. After half an hour stability we started the infusion. From that moment we knew that the compound was inside the rat's brain and an effect can appear, and than you gradually see an effect. Most of the times we hardly saw anything, but a few times we saw a slight decrease. So the D_1 agonist has little effect, it hardly deviates from the start condition, but you have the feeling that it has a slight inhibiting effect. Under these conditions amphetamine has a similar slight effect. You also have the feeling that it suppresses, but only very little. So if any, it seems to work in the same way as a D_1-agonist.

58 By trying a D2-agonist we saw that it did something different, it gave a gradual increase of activity. Therefore D_1 and D_2 agonists seem to act differently. But the effects are hardly noticeable. That is why we induced a situation where the striatum was activated. If you infuse a glutamate agonist in the striatum, you see an immediate and very strong effect, that only lasts for a limited amount of time. Hence, what I then looked for was a relatively low dose to create a more stable activation to use as an activated condition.

59 When the experiment is finished we apply a small amount of current on the electrode to burn a little hole, which will mark the location. Then we sacrifice the rat, and remove its brain. You end up with a whole series of jars with brains in them. Then you plan a day when you will slice up all those brains. With the help of the brain atlas and the marked position of the electrode tip you determine the exact location of your measurements.

8.5 Interpreting the data

In this final section I report on my questions to Dr. W. Timmerman about the interpretation of laboratory data in general, and the published conclusion about her investigation in particular.

60 Sometimes the data you obtain deviates from what you expected, or the outcome of one experiment is very different from the rest. In the latter case it is possible that the probe location was wrong. This would give you a reason to remove these results from your sample. However, if that is not the case you will have an anomaly.

61 If I find an anomaly I check the experiments just before and after in the same series to see if something can be traced from that. Also the experiment has to feel right. For instance, sometimes a signal is hardly noticeable in the noise, and then it already casts some doubt.

62 But if the template was good, the stability was in order, and you still see a deviating response, and it was one deviation in *e.g.* five others, then I just mention it in the results section of an article. One rat was an exception for an unknown reason, so be it. As an average we always repeat an experiment five to six times. You cannot base anything on one observation. Sometimes we follow one experiment, but often it turns out that it is still different. You cannot publish anything based on one experiment.

63 Another influence on your data is the anesthetic. For instance, ketamine is an anesthetic that acts on the glutamate receptor. You do not want that. There are all kinds of arguments to use anesthetic. It is less stressful for the animal. You have more stabile activity compared to animals that are awake. But because the striatum is involved in motor behavior you never can be sure that it does not influence your data. You do not know until you also check it with awake animals.

64 Yet another factor is that in Parkinson research animal models are used that are lesioned with for instance MPTP [see Section 7.2], but that may not reflect the entire or precise pathological situation. So conclusions about the model may not be true for the disease.

65 Another issue is that many effects in experiments with systemic dopaminergic injections are ascribed to the striatum. This is indeed one of the areas where effects can be mediated. But an effect can also be directly induced in the accumbens or the SNR. You could be wrong by claiming that it was the striatum. DA released from dendrites can also be involved. That is another complication. So it is not all that easy to establish the functional role of dopamine.

66 For the manipulations in our research we focused on glutamate and dopamine interactions. But in the back of your mind you know that there are also dopamine-acetylcholine interactions, and all kinds of peptides, and the influence of GABA-ergic neurons. So you leave out a great many to keep a grasp on what you are doing. So if you find things that you can not easily understand, there are many explanations possible. You know you cannot explain everything by just measuring dopamine, glutamate and GABA, there is much more to it.

This concludes my interviews with Dr. Timmerman about her work and experiments up to February 1997. In later tests Dr. Timmerman further experimented with different setups, such as beginning with a glutamate agonist infusion, followed with a glutamate agonist infusion in combination with DA-agonist. About this work she and her coworkers published the following conclusion:

67 "To gain insight into the role of striatal dopamine in basal ganglia functioning, dopaminergic drugs alone, and in combination with the glutamate receptor agonist kainic acid were infused in the lateral striatum via a microdialysis probe, while single-unit recordings of substantia nigra reticulata neurons were made in chloral hydrate-anaesthetized rats. Striatal infusion of dopaminergic drugs did not significantly affect the firing rate of substantia nigra reticulata neurons, which was related to the low activity of striatal cells under basal conditions, illustrated by the lack of effect of striatal infusion of TTX on substantia nigra reticulata activity. Under glutamate-stimulated conditions, striatal infusion of *d*-amphetamine potentiated the inhibition of substantia nigra reticulata neurons induced by striatal kainic acid. Thus, under stimulated but not basal conditions, the modulatory role of dopamine in the striatum could be demonstrated. Dopamine potentiated the inhibitory effect of striatal kainic acid on the firing rate of the basal ganglia output neurons." (W. Timmerman, F. Westerhof, T. van der Wal and B. Westerink, 1998)

8.6 Conclusion

The specific question for this chapter was: how are theory and experiments used in drug research for Parkinson's disease, in practice? I tried to present an image of a practice in neuropharmacology by interviewing two scientists about their specific work in investigating new drugs, exploring the functions of part of the brain, testing a model of those functions and interpreting the data.

Overall, neuropharmacologic research can be characterized as searching, understanding and testing a way to make the characteristics of a pathological systems resemble a healthy situation. Experiments are used to chart both situations, and to try to bring one situation closer to the other by drug manipulations. In the next chapter I will analyze the specific problems addressed in the practice, as described in this chapter, in detail.

* *

Chapter 9

Discovery

9.1 Introduction

In Chapter 8 I reported on my own epistemological experiment, where I observed and inquired about a scientific practice, the process of discovery in neuropharmacology. Chapter 7 reported on a part of the theory that is used and developed in that practice. In this final chapter I analyze both the theory and practice, using the concepts from my theoretical discussion of discovery in Part II. The particular question that is answered in this chapter is: what is the rational use of theory and experiment in neuropharmacology? For my description of discovery in neuropharmacology I will pursue answers to the three specific questions of this thesis, *i.e.* 1) what is the structure of a scientific theory?; 2) what is the process of scientific reasoning?; and 3) what is the route between theory and experiment?

In answering these questions in this chapter I combine the theoretical approaches of logic as introduced in Chapter 4, and cognitive science, as discussed in Chapter 5. My main goal is to describe the practice of neuropharmacology. I will use the problem solving concepts from cognitive psychology to describe steps in the *process* of discovery, while I use the concepts of the logical approach to describe the *products* of that process. My aim is not to explain the particular directions of the search process that is described, by extracting and representing implicit knowledge as production rules. Those rules are dependent on the personal experiences of researchers and learned in a particular practice, as argued in Sections 5.7 and 5.8.

To analyze the structure of the DA theory I will first, in section 9.2, introduce a logical approach to represent the structure of theories in general, and dynamical systems in particular. Then, in section 9.3, I formally represent the theory of the basal ganglia as a qualitative differential equation, to answer the first question of this thesis for the case study. Before going into the second question, the third question is addressed in section 9.4, where I describe the route between theory and experiment in the problems faced in the practice of neuropharmacology. In section 9.5 I go into the process of reasoning in explanation, prediction and design. I will also discuss how a description of that process could be applied in that practice. Finally, in section 9.6 I end with a general conclusion, discussing the consequences of my observations and analysis of the case for the theory about discovery as discussed in Part II.

9.2 Models

The first question I will address is, how to understand the structure of the DA theory of Parkinson's disease. And secondly, how does it explain the effect of known treatments. In this section I will introduce a model theoretic approach to the structure of theories.

The structuralist approach in the philosophy of science characterizes a theory by its models, conceived as structures. (Th.A.F. Kuipers, 2000). A structure, in this context, is usually represented as an ordered set of variables, functions and constants. A structure is called a model of a theory if the theory, seen as a proposition about that structure, is true.

The core of a theory consists of a set of models M which is a subset of all conceptually possible models M_P given the vocabulary of the theory. The difference between M_P and M are all models that the theory excludes and is called the empirical content of a theory. It contains all the potential falsifiers of the theory. Given a domain D of application of the theory it is assumed that there is a subset of M_P that are the empirically possible models of that domain. A weak empirical claim states that all empirically possible models are models of the theory, a strong claim also asserts that they are equal.

For my exposition I will characterize a theory by its vocabulary of variables, the quantity spaces of those variables (a quantity space of a variable defines the range and type of values of a variable), and constraints on the values of those variables, given that they represent together the set of possible models and models of the theory. I will further make a distinction between a theory T, which is basically a set of definitions, and a hypothesis H which is a statement that asserts that the properties of phenomena in domain D can be characterized by the vocabulary V and by the models of theory T.

Definition 1 *Theory.* The ordered set $\langle V, Q, C\rangle$ of variables V, quantity spaces Q and constraints C represents a theory. The theory determines an ordered set $\langle M_P, M_T\rangle$ that contains the conceptually possible models M_P, given V and Q, and the models of the theory M_T, given the constraints C on V.

Definition 2 *Hypothesis.* The ordered set $\langle V, Q, C, D\rangle$ represents a hypothesis where a theory is applied to a domain D. The hypothesis determines the ordered set $\langle M_P, M_T, M_E\rangle$, that contains the conceptually possible models M_P of a domain D given possible descriptions by variables V and quantity spaces Q; the models M_T of the theory of the domain given constraints C on V; and the empirically possible models M_E of the phenomena of domain D. The hypothesis asserts that the set of empirically possible models M_E is a subset of, or equal to, the set of models M_T of the theory.

A model of a phenomenon in a domain is a structure that represents certain aspects of that phenomenon in terms of a set of interpreted variables with particular quantities. The structures that are possible according to the constraints C from a theory are called the models M_T of that theory. The conceptually possible models M_P is the set of all the models that are possible if you combine all possible variables from V with all their possible quantities from Q.

The relation between the conceptually possible models M_P, the models of the domain M_E and the models M_T of a theory in a hypothesis can be graphically represented as in Figure 9.1. The different intersections represent subsets of structures that constitute either a success, an anomaly or a problem for the theory. The goal of explanation is to find a hypothesis, such that a better hypothesis has less problems (subset 1) or anomalies (subset 3) then a competitor (Th. Kuipers, 1992, p.303).

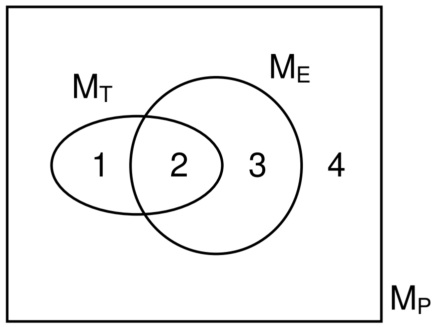

Figure 9.1: Models M_T of a hypothesis and empirically possible models M_E of the phenomena of a domain, both part of the conceptually possible models M_P

Subset	M_T	M_E	
1	1	0	Explanatory problem
2	1	1	Explanatory success, confirming instance
3	0	1	Empirical anomaly, counter example
4	0	0	Explanatory success

Table 9.1: Subsets of conceptually possible models M_P of a domain

To understand the theory of Parkinson's disease we can understand it to be a hypothesis about the dynamical behavior of the brain. The theory asserts what kind of states and behaviors are possible. The set V and Q describe the known structural properties of the brain, and the constraints in C describe the assumed functional relations between those properties. A variable x of a structure is related to variable y if there is a functional constraint in C, such that $y = f(x)$.

Disease and intervention

To understand the research problems in pharmacology we need to extend our vocabulary. Pharmaceutical research is not only interested in how to explain observations of a pathological biological system. It also aims to know how to treat it, and why a treatment works. For this we can introduce two extra subsets of M_P, the models of a biological system that is influenced by a (drug) intervention, M_I, and the models of phenomena that we wish to cause, the set M_W, see Figure 7.3.

Given a set of conceptually possible models of the behavior of a biological system a set of drug interventions can be assumed to cause behaviors represented by the set M_I, while the set M_W represents the set of wished for behaviors. Let M_E represent the empirically possible behaviors of a living organism with a given biological structure. Hence if the assumptions are correct M_I should be a subset of M_E.

In Figure 9.2 subset 1 denotes an undesired behavior that is not treated by known interventions. Subset 2 contains unsuccessfully treated system behavior and unwanted side effects of a partially successful drug treatment, while subset 3 denotes behavior that is successfully treated. Subset 4 may be equal to health, given that W denotes health. Subset 5 can contain a behavior that is not possible given the biological structure of the organism, but can still be desired. Subset 6 equals the periphery of both possibility and interest.

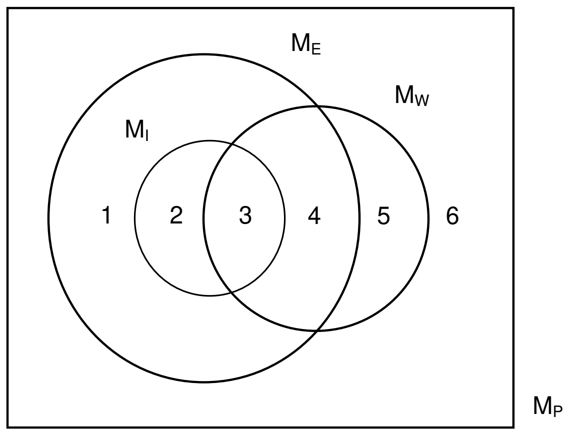

Figure 9.2: Empirically possible models M_E of a biological system, wished for models M_W, and models M_I of a system that is influenced by an intervention, all part of conceptually possible models M_P of a biological system

Subset	M_E	M_I	M_W	
1	1	0	0	Disease, untreated by known interventions
2	1	1	0	Disease, treated with side effects
3	1	1	1	Successfully treated
4	1	0	1	Health
5	0	0	1	Desired, but not empirically possible
6	0	0	0	periphery of interest and possibility

Table 9.2: Subsets of conceptually possible models M_P of a biological system

These three sets define the main goals of neuropharmacology. It is a goal to describe and explain M_E, what kinds of values of variables describing the brain and behavior of the organism are empirically possible, and why? It is also a goal to determine what kinds of states and behaviors M_W constitute health, or are desired for other reasons. And finally what kind of drug or other medical interventions cause those desired behaviors.

Dynamical systems

In neurobiology the function of the brain is described and explained as a complex dynamical system. In physics, the most powerful tool to model a dynamical system is by making use of differential equations. Variables represent properties of the system whose values can change over time. By defining the specific relations between those variables those values can be predicted, given an initial state of the system.

Empirical studies of both the brain and behavior in Parkinson research results in many quantitative data, correlating variables of the activation frequency of nuclei and neural pathways and local concentrations of different kinds of neurotransmitters. Yet those relations are not sufficiently known to define them as a quantitative equation. The relation is only known qualitatively. Many results of empirical studies of the brain amount to conclusions such as *e.g.* if the value of this variable changes in this direction, the change of the value of that variable in that direction is statistically significant. In this way the theory that explains Parkinson's disease can explain why the activation of the thalamus decreases, when the concentration of DA in the striatum significantly decreases.

While these results are insufficient to define a model with the aid of an ordinary differential equation, they can be represented by a more abstract qualitative differential equation (QDE), *cf.* (B. Kuipers, 1994). A QDE can be defined as follows:

Definition 3 *Qualitative differential equation.* A Qualitative differential equation (QDE) represented by the ordered set $\langle V, Q, C\rangle$ is an abstraction of an ordinary differential equation (ODE):

$$dV/dt = C(V)$$

where V is a set of variables each of which is a *reasonable function* over time, whose values are described in a finite set of qualitative landmark values belonging to Q, and C is a set of constraints between those variables.

Definition 4 *Reasonable function $v \in V$.* Given an interval of the set of real numbers extended with ∞ and $-\infty$, $[a,b] \subseteq \Re^*$, the function $v: [a, b] \rightarrow \Re^*$ is a reasonable function over [a, b] if

1. v is continuous over [a, b]
2. v is continuously differentiable over (a, b)
3. v has only finitely many critical points in any bounded interval,
4. the one-sided limits $\lim_{t \rightarrow a} v'(t)$ and $\lim_{t \rightarrow b} v'(t)$ exist in $\Re^*$ and are defined as to be equal to v'(a) and v'(b), respectively.

Definition 5 *Quantity space $q \in Q$.* A quantity space *q* is a finite, totally ordered set of landmark symbols such as $-\infty < l_1 < ... < 0 < ... < l_k < \infty$ that describe qualitatively important distinctions for a variable.

Important distinctions described by landmarks are values of variables that change in time and become steady or start to increase or decrease at a certain time-point. The qualitative value of a variable v at time-point t is expressed by a landmark from its quantity space and a direction of change.

Definition 6 *Qualitative value at a time-point.* The qualitative value QV(v, t) of a variable v(t) with respect to a quantity space $q = \langle l_1, ...l_n \rangle$ is defined by the tuple ⟨qmag, qdir⟩ where,

qmag =	l_i	if $v(t) = l_i$,
	(l_i, l_{i+1})	if $v(t) \in (l_i, l_{i+1})$
qdir =	inc	if $v'(t) > 0$
	std	if $v'(t) = 0$
	dec	if $v'(t) < 0$

In a QDE the possible values of the variables in V are constrained by constraints in C. The constraints in C can consist of constraints corresponding to additions, multiplications, negations, derivatives, and incompletely known functions specified only as being part of a monotonicity class.

The last category is relevant for our case. We can know about a function f between two variables $v_1(t)$ and $v_2(t)$, $v_1(t) = f(v_2(t))$, that it belongs to either M^+, the class of monotonically increasing functions, or M^-, the class of monotonically decreasing functions. That is, for every $f \in M^+$, $f' > 0$, and for every $f \in M^-$, $f' < 0$ over the domain of the function. These classes can be generalized to multivariate functions so that *e.g.* $M^{+\,-}$ is the class of functions $v_1(t) = f(v_2(t), v_3(t))$, such that $\partial f/\partial v_2 > 0$ and $\partial f/\partial v_3 < 0$.

The constraints C in a QDE define which qualitative states and behaviors are possible. So C amounts to a theory about a system. We can define the qualitative state of a dynamical system at a distinguished time-point, or on an interval between two distinguished time-points.

Definition 7 *Qualitative state.* The qualitative state of a dynamical system described by m variables V at time-point t_i is an ordered set of individual qualitative values at a certain time-point or time interval from $t_{i,}$ to t_{i+1}:

$$QS(V, t_i) = \langle QV(v_1, t_i), ..., QV(v_m, t_i)\rangle$$
$$QS(V, t_{i,} t_{i+1}) = \langle QV(v_1, t_{i,} t_{i+1}), ..., QV(v_m, t_{i,} t_{i+1})\rangle$$

The qualitative behavior of a dynamical system can now be defined as an ordered set of qualitative states:

Definition 8 *Qualitative behavior.* The qualitative behavior of a dynamical system with variables V on time interval $[t_0 < \ldots < t_n]$ is a sequence of qualitative states:

$$QB(V) = \langle QS(V, t_0), QS(V, t_{0,} t_1), QS(V, t_1), ... , QS(V, t_n)\rangle$$

The possible states and behaviors of a system can be seen as models of the differential equation. Benjamin Kuipers developed a computer program called QSIM that can generate those models. It takes as input a QDE and an initial qualitative state description and produces a tree of possible state sequences. This can be seen as:

$$\mathrm{QSIM}(\langle V,Q,C\rangle, \mathrm{QS}(t_0)) = M$$

such that M is an ordered set ⟨S, B⟩, where S is a set of all possible qualitative states and B is a set of all possible qualitative behaviors, *i.e.* totally ordered sets of qualitative states, consistent with C (*cf.* Schultz and B. Kuipers, 1994).

It can be proved that given an ordinary differential equation (ODE) and its QDE abstraction, all abstractions of genuine behaviors of the ODE are generated by QSIM, but also some behaviors that are not an abstraction of a genuine ODE behavior. It can predict spurious behaviors, not predicted by a numerical solution of the ODE. It remains an open research problem whether qualitative solutions can be made complete or are inherently incomplete. But to put this problem in perspective, numeric algorithms also may produce non-sensical solutions due to round-off errors and careless simulation around singular points (De Jong, 1998).

In the next section I will use the QDE representation to explicate the structure of the dopamine theory of Parkinson's disease, and how it explains the function of known treatments.

9.3 Theory

Neurobiologists study the processes of the brain, *e.g.* by recording values of activation frequencies and concentrations of neurotransmitters in different locations of the brains of guinea pigs, Wistar rats, or monkeys. When the values of two variables v_1 and v_2 are consistent with a monotonic function in all trials of an experiment, a correlation could be proposed.

This is a simple style of descriptive induction, they are monotonically related in the sample, so they are monotonically related in all brains, of the sample organism or even in the human brain. It becomes an explanation if it is hypothesized what processes make it so that the variables act in that way.

In Parkinson research it is observed that the increase of symptoms is correlated with a substantial decrease of the availability of the neurotransmitter DA, which is due to a decay of the substantia nigra pars compacta (SNC). The model of the basal ganglia aims to explain why the decrease of DA can lead to these symptoms, by explaining why the activation of the SNR increases as a result of this decrease.

I will formally reconstruct this explanation by first representing the model of the basal ganglia as a qualitative differential equation. This equation serves as a hypothesis from which it can be deduced that given a decrease of DA, an increase of the SNR activation is a consequence.

I will also show how the activity of known treatments can be explained as well. This elaborate reconstruction will serve my analysis in section 9.4 about how the basal ganglia hypothesis itself is tested experimentally. The process of drug design will be discussed in section 9.5, where I will argue how such explicit models can be used to infer possible new interventions.

Theory of the basal ganglia

The basal ganglia theory is a qualitative theory about a dynamical system, so we can represent it as a QDE. We defined a QDE as a tuple ⟨V, Q, C⟩, where V is a set of variables which are reasonable functions over time, Q is a set of quantity spaces for those variables, and C is a set of constraints. In the basal ganglia theory there are two basic variables describing firing rate (*f*) of nerve cells in a cell group, nuclei or pathway, and the amount (*a*) of a particular neurotransmitter released in the vicinity of a cell group, nuclei or neural pathway. The relation d/dt y = M^+ x is used to state that the change of values of y over time is monotonically related to the change of value of x. It is a matter of debate whether this relation represents a causal direction from x to y, for discussion see Iwasaki & Simon (1994).

I will represent the model of the basal ganglia as depicted in Figure 3.1, which was used by dr. Timmerman (1992). While this model could be further extended to include other influences, such as those of substance P and enkephalin, as depicted in Figure 7.4, the simpler model suffices for my analysis of the observed practice. The notation x-to-y in the cell groups denotes the neural pathway from cell group x to cell group y. I further abbreviate SNR/Gpi to SNR. So, we can define the basal ganglia theory as follows:

Definition 9 *Basal ganglia theory.* T_{BG} : ⟨V, Q, C⟩ is a QDE such that:

1. Variables in V

- Cell groups G, containing nuclei and neural pathways

  G: {striatum, GPe, STN, SNR, thalamus, brainstem, cortex-to-striatum, SNC-to-striatum, striatum-D1-to-SNR, striatum-D2-to-GPe, GPe-to-SNR, GPe-to-STN, STN-to-SNR, SNR-to-thalamus, SNR-to-brainstem}

- Set of neurotransmitters N: {Glu, DA, GABA}

- The firing rate *f*(g) of cell group g is a value of quantity space F

  f: G $\rightarrow$ F

- Amount *a*(n, g) of neurotransmitter n in cell group g is a value of A

  a: N $\times$ G $\rightarrow$ A

2. Quantity spaces in Q

- Boundaries of firing rates F: {0, MAX}

- Boundaries of amounts A: {0, MAX}

3. Constraints in C

- Firing rate of nuclei in the basal ganglia

C.1 d/dt *f*(striatum) = M^{+} *a*(Glu, striatum)
C.2 d/dt *f*(GPe) = M^{-} *a*(GABA, GPe)
C.3 d/dt *f*(STN) = M^{-} *a*(GABA, STN)
C.4 d/dt *f*(SNR) = M^{-+}(*a*(GABA, SNR), *a*(Glu, SNR))
C.5 d/dt *f*(thalamus) = M^{-} *a*(GABA, thalamus)
C.6 d/dt *f*(brainstem) = M^{-} *a*(GABA, brainstem)

- Firing rate of neural pathways between nuclei

C.7 d/dt *f*(cortex-to-striatum) = M^{+} *f*(cortex)
C.8 d/dt *f*(SNC-to-striatum) = M^{+} *f*(SNC)
C.9 d/dt *f*(striatum-D1-to-SNR/GPi) = M^{++}(*f*(striatum), *a*(DA, striatum))
C.10 d/dt *f*(striatum-D2-to-GPe) = M^{+-} (*f*(striatum), *a*(DA, striatum))
C.11 d/dt *f*(GPe-to-SNR) = M^{+} *f*(GPe)
C.12 d/dt *f*(GPe -to-STN) = M^{+} *f*(GPe)
C.13 d/dt *f*(STN-to-SNR) = M^{+} *f*(STN)
C.14 d/dt *f*(SNR-to-thalamus) = M^{+} *f*(SNR)
C.15 d/dt *f*(SNR-to-brainstem) = M^{+} *f*(SNR)

- Amount of released neurotransmitters in nuclei

C.16 d/dt *a*(DA, striatum) = M^{+} *f*(SNC-to-striatum)
C.17 d/dt *a*(Glu, striatum) = M^{+} *f*(cortex-to-striatum)
C.18 d/dt *a*(GABA, GPe) = M^{+} *f*(striatum-D2-to-GPe)
C.19 d/dt *a*(GABA, STN) = M^{+} *f*(GPE-to-STN)
C.20 d/dt *a*(GABA, SNR) = M^{++}(*f*(striatum-D1-to-SNR), *f*(GPe-to-SNR))
C.21 d/dt *a*(Glu, SNR) = M^{+} *f*(STN-to-SNR)
C.22 d/dt *a*(GABA, thalamus) = M^{+} *f*(SNR-to-thalamus)
C.23 d/dt *a*(GABA, brainstem) = M^{+} *f*(SNR-to-brainstem)

- Metabolism of dopamine

C.24 d/dt *a*(DA, x) = *a*(L-dopa, x) × Enzyme-ratio
C.25 Enzyme-ratio = *a*(AADC, x) / *a*(MAO-B, x)

For brevity I include the assumptions about the metabolism of dopamine as part of the theory of the basal ganglia. The availability of dopamine outside the dopaminergic cell terminal is dependent on the activation of the cell by the neural pathway from the SNC. But DA can only be released by the vesicles of the terminal if the precursor L-dopa and the enzyme AADC is available. The enzyme MAO-B breaks down the excess of dopamine to DOPAC, see Figure 7.3.

Explanation of Parkinson's Disease

The theory of the basal ganglia can be applied to explain observations in Parkinson's disease research. The hypothesis of the basal ganglia states that the empirically possible states E of the basal ganglia, given the empirical study of the basal ganglia D, are part of the theoretically possible states M.

Definition 10 *Basal ganglia hypothesis.* $H_{BG} : \langle V, Q, C, D\rangle$ represents a hypothesis about the basal ganglia brain structure where V, Q, C are part of the T_{BG} and D is the set of instances of the basal ganglia, the domain of application of the theory.

We saw that the symptoms of Parkinson's disease are assumed to be caused by an increase of activation of the SNR, which on its turn is explained by a steep decrease of DA in the striatum due to the decay of dopaminergic nerve cells. One question in this chain, how the observed decrease of DA causes the assumed increase of SNR activation, is explained by the theory about the basal ganglia. I will show how this proposition can be deduced from the basal ganglia theory by programs like QSIM. In this proposition and proof I will reduce the values of the variables to just their qualitative direction, abstracting from time and qualitative magnitude.

From d/dt $y = f(x)$ where $f \in M^{+}$ we know that x and y both increase or decrease together, while if $f \in M^{-}$, y increases when x decreases, and vice versa. If $dz/dt = f(x, y)$ and $f \in M^{++}$, the direction of change of z is unknown if x increases and y decreases, since we do not now their magnitude, *cf.* Table 9.3. This is similar for $f \in M^{+-}$, when both variables increase or decrease in value

y \ x	inc	std	dec
inc	inc	inc	?
std	inc	std	dec
dec	?	dec	dec

Table 9.3: Derivative values for z if $dz/dt = f(x,y)$ and $f \in M^{++}$

As background assumptions we assume that the amount of dopamine in the striatum decreases and the firing rate of the striatum is steady. I will use the notation v = qdir as shorthand for $\exists y \exists t\ QV(v, t) = \langle y, qdir\rangle$.

Theorem 1 $H_{BG} \cup B$: $\{a(\text{DA, striatum}) = \text{dec}, f(\text{striatum}) = \text{std}\} \vdash$
P: $\{f(\text{SNR}) = \text{inc}\}$

Proof: As a proof I deduce the conclusion P from the premises B by applying the constraints C from the basal ganglia hypothesis H_{BG}.

a(DA, striatum) = dec $\wedge$ f(striatum) = std
$\Rightarrow$ f(striatum-D1-to-SNR) = dec $\wedge$ f(striatum-D2-to-GPe) = inc (C.9, C.10)

f(striatum-D2-to-GPe) = inc
$\Rightarrow$ a(GABA, GPe) = inc (C.18)
$\Rightarrow$ f(GPe) = dec (C.2)

$\Rightarrow f(\text{GPe-to-SNR}) = \text{dec} \wedge f(\text{GPe -to-STN}) = \text{dec}$ (C.11, C.12)

$f(\text{GPe -to-STN}) = \text{dec}$
$\Rightarrow a(\text{GABA, STN}) = \text{dec}$ (C.19)
$\Rightarrow f(\text{STN}) = \text{inc}$ (C.3)
$\Rightarrow f(\text{STN-to-SNR}) = \text{inc}$ (C.13)
$\Rightarrow a(\text{Glu, SNR}) = \text{inc}$ (C.21)

$f(\text{GPe -to-SNR}) = \text{dec} \wedge f(\text{striatum-D1-to-SNR}) = \text{dec}$
$\Rightarrow a(\text{GABA, SNR}) = \text{dec}$ (C.20)

$a(\text{Glu, SNR}) = \text{inc} \wedge a(\text{GABA, SNR}) = \text{dec}$
$\Rightarrow f(\text{SNR}) = \text{inc}$ (C.4) (Q.E.D)

Explanation of known treatments

I will first introduce a new set in my terminology. Next to a hypothesis H, background assumptions B, and propositions P that are explained or need to be explained, we also have a set of interventions I. This set contains propositions that describe a property of the world, usually a value of a particular variable, that can be set by a manipulation. All consequences of that manipulation hold for all the structures in the set M_I.

A theory can explain why a particular intervention has a particular consequence. With H_{BG} we have a hypothesis that explains the symptoms of Parkinson's disease by linking them to the observed decrease of DA. The hypothesis also explains the function of metabolites like L-dopa, MAO-B and AADC. These metabolites can serve as an artificial intervention by changing their amount with the aid of a drug. Parkinson drugs all serve to increase the amount of dopamine, which according to the theory would decrease the overactivation of the SNR, reducing the behavioral symptoms. In the theorems below I demonstrate how the basal ganglia hypothesis explains the activity of known drug interventions for Parkinson's disease. All these drugs aim to influence the amount of dopamine, so I first pose the following theorem:

Theorem 2 $H_{BG} \cup$ B: $\{f(\text{striatum}) = \text{std}\} \vdash$
P: $\{a(\text{DA, striatum}) = \text{inc} \rightarrow f(\text{SNR}) = \text{dec}\}$

From H_{BG} it can be deduced in theorem 2 that an increase of DA implies a decrease of the firing rate of the SNR output nuclei of the basal ganglia. The proof follows similar lines as the proof of theorem 1.

Theorem 3 states that an increase of L-dopa in the striatum will increase DA in the striatum, which is a consequence of C.24, and given that the enzyme ratio does not increase.

Theorem 3 $H_{BG} \cup$ I: $\{a(\text{L-dopa, striatum}) = \text{inc}\} \vdash$ P: $\{a(\text{DA, striatum}) = \text{inc}\}$

But to increase L-dopa by a drug intervention, which is taken up in the bloodstream, means that L-dopa is increased in the whole body, causing side effects. A decrease of the amount of AADC in the periphery by also administering an inhibitor that can not cross the blood brain barrier, will cause DA to increase in the brain, but to be relatively steady in the periphery. This theorem (4) is a consequence of C.24 and C.25, given the assumption that the amount of MAO-B does not increase in the periphery.

Theorem 4 $H_{BG} \cup$ I: {a(L-dopa, body) = inc, a(AADC, periphery) = dec} $\vdash$
P: {a(DA, striatum) = inc, a(DA, periphery) = ?}

By C.25 and C. 25 one can also prove theorem 5, which states that decreasing the enzyme that breaks up DA will increase he amount of DA, assuming that both the amount of AADC and L-dopa in the striatum do not increase:

Theorem 5 $H_{BG} \cup$ I: {a(MAO-B, striatum) = dec} $\vdash$ P: {a(DA, striatum) = inc}

The function and activity of these treatments can be explained by the theory of the basal ganglia, but another question is if the hypothesis is true. That is, are all the states that are empirically possible also states allowed by the theory?

In Section 8.4 we saw how experiments showed that the background assumption in theorem 2 about the activation of the striatum turned out to be incorrect, by testing the predicted effect of selective dopamine receptor agonists. In the next section I will go into that problem.

9.4 Practice

In the second part of this thesis I discussed several ways of understanding rationality in the process of scientific discovery. In this discussion it was assumed that the main goal of scientific discovery is to gain knowledge about natural phenomena. In order to do so I made a distinction between five basic types of tasks as problems with different sub-goals: observation, description, explanation, prediction and intervention. These tasks, repeatedly executed in that order, could lead to an increasingly better knowledge of the natural world.

In this section I extensively analyze the actual problems and tasks that are tackled in the practice of neuropharmacology that I observed. To describe a problem I will make use of the distinctions made in Chapter 5. That is, pursuing a problem is characterized by the following constituents:

Problem

Start:	propositions about the initial situation
Goal:	the conditions for a problem to be solved
Result:	the propositions describing the result
Process:	the kind of action that is used to pursue the goal.

To abstractly distinguish different contents of propositions I will make use of different sets of propositions that describe:

Propositions

O:	observations
I:	interventions
H:	hypotheses
W:	wished for properties
P:	predictions

A question-mark after the name of a set, *e.g.* H?, will designate the set or property that is the goal of the problem. A star after the name of a set, *e.g.* H*, or a star in a set, *e.g.* H: {*}, will mean that the set contains propositions, for which the truth-value is unknown, that describe new information relative to the initial situation. Processes that can lead to achieving a goal are distinguished as:

Processes

Intervention:	manipulating a property of a natural process
Observation:	observing a property of a natural process
Description:	describing a property of a natural process
Explanation:	finding an explanation for the initial situation
Prediction:	deducing a consequence of a initial situation
Design:	creating a property given a specification

Design as a process is added because it is needed to describe some problems of neuropharmacology involving wished for properties of a drug or a treatment. I will make a further distinction between a focused and broad kind of process, meaning that the activity is directed to a small or large set of properties.

In Chapter 5 these terms where used to designate particular problems and processes with a particular initial and goal situation, see table 5.10. I will now use them to describe (parts of) larger scale problems. In the next section I will go deeper into the process of reasoning and compare the theoretical archetypes with my observations.

In describing a problem I will first model the observed examples from the practice. I will loosely follow the order of the report of my interviews in Chapter 8. A problem never comes by itself, so I will describe different situations where the result of a particular problem leads to new problems that are addressed in that situation.

In Chapter 8 I reported on several situations that lead to different kinds of discoveries: new functions of known drugs are discovered in the clinic and further investigated in the lab; it is investigated what a desired function of a new drug should be; given that wished for function new drugs are designed, searched, predicted, created and tested; new drugs are tested and investigated in the clinic; they are also used to explore biological systems in the lab; given that exploration, new treatments are designed and tested; and the function of a drug is explored and explained. Below I analyze the structure of those problems. The specific problem of exploring and explaining the function of DA is analyzed in detail. I summarize and generalize the examples in Table 9.4 at the end of this section.

Discovering a new drug effect in the clinic

In Section 8.2 It is reported that chlorpromazine was specifically administered to treat particular diseases in the clinic (*cf.* Sec. 8.2, Par. 2). It was noted that another aspect than the one that was to be treated was influenced as well. This was followed by lab investigations that aimed to first observe and to explain the drug effect on another level. So we can distinguish a sequence of three kinds of problems: clinical drug treatment; lab investigation of the observed drug effect; explanation of the observed drug effect. This practice can be described as follows (*cf.* Table 9.1a):

Clinical drug treatment

Start: O: {particular disease}
Goal: Clinic I: {chlorpromazine} → W: {treatment}
Result: I → Clinic O*: {parkinsonism}
Process: Focused intervention, broad observation

Lab investigation of observed drug effect

Start: Clinic I: {chlorpromazine } → O
Goal: Lab I: {chlorpromazine} → O?
Result: Lab O*: {amount of DA release}
Process: Focused intervention, focused observation

Explanation of observed drug effect

Goal: H? ⊨ I: {chlorpromazine} → O*: {amount of DA}
Result: H: {chlorpromazine = DA-antagonist,
DA is related to Parkinson' disease}
Process: Focused explanation

Searching a desired drug effect

In the above situation a potential treatment for a disease was discovered while it was initially not the goal of the activity that led to the discovery to study the function of that drug for that disease. Given that a desired effect was observed it was further investigated what the specific effect was on a lower biological level (*cf.* Sec. 8.2, Par.2).

When it was discovered that dopamine was related to Parkinson's disease it was investigated what the specific effect of dopamine is, to determine what a drug that should act as dopamine should do. So part of the result of the investigation is a profile for a new drug. The activity of a new drug such as a specific DA receptor agonist can then be compared to the desired activity of dopamine. The problem goes as follows (*cf.* Table 9.1b):

Lab exploration of wished drug effects

Start: I: {in vivo/ vitro DA}
Goal: I → O?: {receptor activity?}
Result: I → W*: {amount of C-Amp release}
Process: Focused intervention, focused observation

Lab testing for wished drug effect

Start: I: {in vivo/ vitro, specific DA-agonists}
Goal: $I \rightarrow W?$: {receptor activity?}
Result: $I \rightarrow O^*$: {amount of C-Amp release}
Process: Focused intervention, focused observation

Designing a new drug

In Groningen professor Horn aimed to design a variant of dopamine that had a similar activity and metabolism as dopamine itself, but had also effects that made it more useful as a drug, such as specific receptor selectivity and lipophilicity (*cf.* Sec. 8.2, Par. 4-5). The suggestion for variants that Prof. Horn considered where based on his experience and knowledge of the field (Par. 6). While that knowledge may not always be explicit, it could still imply the suggestions. Success of a suggestion is hard to predict given the partly uncertain and incomplete nature of knowledge at a scientific frontier.

Professor Horn used his knowledge to design variants of ADTN. He could explain how a propyl and hydroxy group could respectively aid the lipophilicity and metabolism of the variant. So we can describe the process as designing, testing and explaining a new drug (*cf.* Table 9.1c):

Rational drug design

Start: I: {ADTN}
Goal: $H \models I? \rightarrow W$: {activity, lipophilicity, metabolism, selectivity}
Result: I*: {ADTN-variant}
Process: Focused design

Lab testing for wished drug effect

Start: I: {ADTN-variant}
Goal: $I \rightarrow W?$: {activity?, lipophilicity?, metabolism?, selectivity?}
Result: $I \rightarrow O^*$
Process: Focused intervention, focused observation

Explanation of drug effect

Start: I: {ADTN-variant} $\rightarrow$ O: {lipophilicity, metabolism}
Goal: $H? \models I \rightarrow O$
Result: H: {propyl group $\rightarrow$ lipophilicity, hydroxy group $\rightarrow$ metabolism }
Process: Focused explanation

Searching a new drug effect in a drug library

For a pharmaceutical company the results of the process designing new drugs leads to a library of novel compounds that are created with a specific goal, a given set of criteria (*cf.* Sec. 8.2, Par. 8). Often these criteria include the selectivity for a particular known receptor. A new drug treatment can be discovered by testing those drugs on other receptors by trial and error. In this process the drugs are given, and only massively tested on one criterion. A compound that is found to be active can be the lead for a new drug to target the new receptor, (*cf.* Table 9.1d).

Lab testing for wished drug effect

Start:	I: {in vitro, all compounds of company on new receptor}
Goal:	I $\rightarrow$ W?: {receptor activity?}
Result:	I $\rightarrow$ O*: {amount of C-Amp release}
Process:	Broad intervention, focused observation

Searching a new drug by combinatorial chemistry

A still broader approach is taken when a drug lead is not specifically varied, based on *fingerspizengefühl* and personal experience, but by techniques from combinatorial chemistry (*cf.* Sec. 8.2, Par. 9). In this approach many variants are created at once. In this process the combinations are made and massively tested for activity. If activity is measured, the responsible variant is retrieved and further explored for its structure, (*cf.* Table 9.1e):

Combinatorial drug design

Start:	I: {drug lead}
Goal:	I? $\rightarrow$ W
Result:	I*: {many variants by combinatorial chemistry}
Process:	Broad design

Lab testing for wished drug effect

Start:	I: {in vitro, all variants on a receptor}
Goal:	I $\rightarrow$ W?: {receptor activity?}
Result:	I $\rightarrow$ O*: {C-Amp release}
Process:	Broad intervention, focused observation

Lab exploration of drug effect

Start:	I: {retrieved drug}$\rightarrow$ O: {high receptor activity}
Goal:	O?
Result:	O*: {chemical structure}
Process:	Focused intervention, focused observation

Searching a new drug by computational modeling

While the trial and error approach in combinatorial chemistry is very expensive, the cheaper knowledge based approach by computer modeling is less successful. In this approach one starts with a computer model of the structure of a receptor and a drug (*cf.* Sec. 8.2, Par. 10). The goal is to predict by a simulation how a drug will dock (interact with a receptor), or how the receptor will fold, (*cf.* Table 9.1f).

Computational drug design

Start:	I, H
Goal:	H $\models$ I? $\rightarrow$ W
Result:	H $\models$ I* $\rightarrow$ P*: {drug docking, protein folding}

Testing a new drug

When a promising new drug or drug function is found and explored in the lab, it will leave the lab for further tests. As reported there are three different test phases (*cf.* Sec. 8.2, Par. 11). In the first phase the drug is tested for its possibly toxic effects on a specifically selected group of animals and volunteers. In phase two the focus of intervention changes to selected patients, where therapeutic effects are tested. In phase three this group is further extended, and the drug is used in hospitals and will undergo double blind tests, (*cf.* Table 9.1g).

Clinical testing Phase I

Start: I: {new drug}
Goal: I: {drug on animals, volunteers} → O?: {toxicity?}
Result: I → O*
Process: Focused intervention, broad observation

Clinical testing Phase II

Goal: I: {volunteer patients} → W?: {therapeutic effect?}
Process: Focused intervention, focused observation

Clinical testing Phase III

Goal: I: {double blind, hospitals} → W?: {therapeutic effect?}
Process: Broad intervention, focused observation

Exploring a biological system

Highly selective drugs are also being used to explore and chart biological systems and to find out the function of specific drugs (*cf.* Sec. 8.2, Par. 12). The goal of Dr. Timmerman is to chart a system like the basal ganglia, using a broad range of selective interventions the explore it (*cf.* Sec. 8.3, Par. 17-18, 20).

Lab exploration of a biological system

Start: I: {different selective agonists, antagonists}
Goal: I → O?: {behavior?, local transmitter response?, electric activity?}
Result: I → O*: {stereotypical beh., amounts of transm., firing frequency}
Process: Broad intervention, broad observation

This is achieved by focusing on the effects of particular drugs. This exploration is also undertaken in the case of a pathological system. For Parkinson's disease the function of dopamine in the basal ganglia is being studied (*cf.* Sec. 8.3, Par. 19). (This case will be analyzed in detail further below). The pathological situation is studied in rats whose dopamine cells are lesioned (*cf.* Sec. 8.3, Par. 35).

Lab exploration of a drug effect

Start: I: {normal rat/lesioned rat, local infusion of selective drug}
Goal: I → O?: {behavior?, local transmitter response?, electric activity?}
Result: I → O*: {stereotypical beh., amounts of transm., firing frequency}
Process: Focused intervention, broad observation

Information about the observed difference between a healthy and pathological system can then be used to understand compensation mechanisms and in the design of new treatments (*cf.* Table 9.1h).

Designing a drug treatment

When an observed difference between a healthy and pathological system is explained on a biochemical level, as in the case of Parkinson's disease, this difference can be used to rationally design a treatment. The goal is to intervene in such a way that the difference is minimized. In Parkinson's disease the goal is to restore the function of dopamine to normal (*cf.* Sec. 8.3, Par. 24-26, 34-35).

Rational drug treatment design

Start: H $\models$ O: {depletion of DA $\rightarrow$ Parkinson's disease symptoms}
Goal: H $\models$ I? $\rightarrow$ W: {restored DA function, best effect on symptoms }
Result: I*: {DA-agonists?, selective D1? and/or D2?, NMDA-antagonists?}

In the case of Parkinson's disease it is not clear how best to restore the function of dopamine. Different interventions are designed and tested on their effect on disease symptoms. But all give rise to different (side) effects (*cf.* Sec. 8.3, Par. 27-30).

Clinical testing of a drug treatment

Start: I
Goal: I $\rightarrow$ W?: {effect on symptoms?}
Result: O: {hyper or poor response, effects in brain periphery, loss of effect}

It is also a problem to know what it means to restore a function to normal. One way to tackle this problems is to vary a dose by trial and error to search for a desired response. Yet targeting the dopamine receptor via the bloodstream also induces nausea via the extra stimulation of DA-receptors in the periphery. Hence part of designing a treatment of disease symptoms is designing treatment for the side effects of that treatment (*cf.* Table 9.1i).

Design of a treatment of side effects

Start: H $\models$ I: {DA-agonists in bloodstream} $\rightarrow$ O: {hyper or poor response}
Goal: H $\models$ I? $\rightarrow$ W: {desired response}
Result: I*: {vary doses DA-agonists by trial & error}

Design of a treatment of side effects

Start: H $\models$ I: {DA-agonists in blood stream} $\rightarrow$ O: {nausea}
Goal: H $\models$ I? $\rightarrow$ W: {DA not in periphery}
Result: I*: {peripheral DA blockers}

Exploring a drug effect

It was discovered that acetylcholine also has an effect on Parkinson symptoms. As an explanation of this effect it was proposed that there might exist a brain mechanism that maintains a balance between acetylcholine and dopamine. This explanation was later confirmed by experimental research that discovered that acetylcholine cells respond to D2-agonists (*cf.* Sec. 8.3, Par. 31-33).

Explanation of a drug effect

Start: I: {acetylcholine antagonist} $\rightarrow$
O: {effect on Parkinson's disease symptoms}
Goal: H? $\models$ I $\rightarrow$ O
Result: H: {acetylcholine-dopamine balance}

Lab exploration of drug effect

Start: I: {D2-agonist, acetylcholine cell}
Goal: I $\rightarrow$ O?: {activity?}
Result: I $\rightarrow$ O*: {inhibition acetylcholine cell}

Dopamine is assumed to be related to the activity of the substantia nigra retaculata (SNR) in the basal ganglia (*cf.* Sec. 8.3, Par. 13-16). The model of the basal ganglia implies that dopamine would act as a modulator of GABA activity, inhibiting the activity of the SNR (*cf.* Sec 9.3).

Explanation of a drug effect

Start: I: {low amount of DA} $\rightarrow$ O: {high SNR activity}
Goal: H? $\models$ I $\rightarrow$ O?
Result: H_{BG}: {DA is modulator of GABA activity in SNR}

To test this claim the interaction between dopamine and the SNR was further explored in Groningen by Dr. Timmerman. She set the problem to investigate the effects of dopamine specifically in the striatum (*cf.* Sec. 8.3, Par. 36). I analyze her approach to this problem in detail, following my report of her experiments, and making use of the QDE formalism of Section 9.3. The problems are summarized in Table 9.4j. The general problem goes as follows:

Exploration of a drug effect

Start: $H_{BG} \models I \rightarrow P$
Goal: I: {dopamine in striatum} $\rightarrow$ O?
Result: I $\rightarrow$ O
Process: Focused prediction, focused observation

The model of the basal ganglia H_{BG} predicts that D1-agonists will excitate the direct pathway to the SNR, while D2-agonists will inhibit the indirect pathway (*cf.* Sec. 8.3, Par. 37-40).

Prediction of a drug effect

Start:	H_{BG}: {DA is modulator of GABA activity in SNR} $\models$ I $\rightarrow$ O
Goal:	$H_{BG} \models$ I: {D1-agonist, D2-agonist}$\rightarrow$ P?
Result:	P*: {D1 excitation of direct pathway, D2 inhibition of indirect pathway}

The problem is now to test whether the implications of the model are correct. First Dr. Timmerman explored three different predicted effects. She locally intervened the amounts of glutamate, a D1-agonist and a D2-agonist in the striatum under basal conditions, and observed the activity of the SNR (*cf.* Sec. 8.4, Par. 41-42). The predicted effects can be deduced from the axioms of H_{BG} in section 9.3.

Lab testing of a predicted drug effect (in vivo)

Start:	B: {*f*(striatum) = ⟨0, std⟩}
Goal:	I $\rightarrow$ O?
Process:	Focused intervention of glutamate, D1 and D2 receptors respectively Focused observation of SNR activity
Start:	B $\cup$ $H_{BG} \models$ I: {*a*(glutamate, striatum) = inc} $\rightarrow$ P: {*f*(SNR) = dec}
Result:	I $\rightarrow$ O*: {*f*(SNR) = dec}
Start:	B $\cup$ $H_{BG} \models$ I: {*a*(D1-agonist, striatum) = inc} $\rightarrow$ P: {*f*(SNR) = dec}
Result:	I $\rightarrow$ O*: {*f*(SNR) = dec, slight}
Start:	B $\cup$ $H_{BG} \models$ I: {*a*(D2-agonist, striatum) = inc} $\rightarrow$ P: {*f*(SNR) = dec}
Result:	I $\rightarrow$ O*: {*f*(SNR) = inc, slight}

Glutamate produced the predicted effect. The intervention with the D1-agonist only produced a very slight effect in the predicted direction, and the D2-agonist produced a slight effect against the predicted direction, but both where not significant. When the D1-agonist is tested in vitro a different effect than the one predicted is observed as well (*cf.* Sec. 8.4, Par. 43).

Lab testing of a predicted drug effect (in vitro)

Start:	B $\cup$ $H_{BG} \models$ I: {*a*(D1-agonist, striatum) = inc} $\rightarrow$ P: {*f*(striatum-D1-to-SNR) = dec}
Goal:	I: {*a*(D1, slices striatum) = inc} $\rightarrow$ O?
Result:	I $\rightarrow$ O*: {*a*(striatum-D1-to-SNR) = inc}
Process:	Focused intervention, focused observation

The explanation that was proposed to account for the observation of the slight effect of the selective dopamine agonists was that under basal conditions, there is no GABA activity to modulate (*cf.* Sec. 8.4, Par. 44). So a starting condition with a higher activation of the striatum should show an effect of a dopamine-agonist infusion.

Prediction of a drug effect

Start:	B: $\{f(\text{striatum}) = \langle 0, \text{std}\rangle\}$, H_{BG}
Goal:	$B \cup H_{BG} \models I$: $\{f(\text{striatum}) = \langle +, \text{std}\rangle, a(\text{DA, striatum}) = \text{inc}\} \rightarrow P^*?$
Result:	P^*: $\{f(\text{SNR}) = \text{inc}\}$

The problem now for exploring the effect of the dopamine agonist is that it is important to maintain a steady activity of the striatum. If two variables vary then it is difficult to explain the effect on a third variable by pointing to only one of those two. So the goal is first to find an intervention that causes a desired effect that is needed as an initial condition for the experiment that tests another intervention (*cf.* Sec. 8.4, Par. 45-47).

Rational design of an experimental condition

Start:	B: $\{f(\text{striatum}) = \langle 0, \text{std}\rangle\}$, H_{BG}
Goal:	$B \cup H_{BG} \models I? \rightarrow W$: $\{f(\text{striatum}) = \langle +, \text{std}\rangle\}$
Result:	I^*
Process:	Focused design

Lab exploration of a predicted drug effect

Start:	B: $\{f(\text{striatum}) = \langle 0, \text{std}\rangle\} \cup H_{BG} \models I \rightarrow P$
Goal:	$I \rightarrow O?$
Process:	Focused intervention, focused observation

For instance:

Start:	$B \cup H_{BG} \models I$: $\{a(\text{glu, striatum}) = \langle +, \text{std}\rangle\} \rightarrow P$: $\{f(\text{striatum}) = \text{std}\}$
Result:	$I \rightarrow O^*$: $\{f(\text{striatum}) \neq \text{std}\}$

Start:	$B \cup H_{BG} \models I$: $\{a(\text{glu, cortex}) = \langle +, \text{std}\rangle\} \rightarrow P$: $\{f(\text{striatum}) = \text{std}\}$
Result:	$I \rightarrow O^*$: $\{f(\text{striatum}) \neq \text{std}\}$

Start:	$B \cup H_{BG} \models I$: $\{a(\text{glu, thalamus}) = \langle +, \text{std}\rangle\} \rightarrow P$: $\{f(\text{striatum}) = \text{std}\}$
Result:	$I \rightarrow O^*$: $\{f(\text{striatum}) \neq \text{std}\}$

Given the model it is predicted that the striatum can be activated directly with glutamate in the striatum, or indirectly with glutamate in the cortex or thalamus (*cf.* Sec. 8.4, Par. 48). However, when tested it turns out that it is difficult to maintain a steady activation. This can be attributed to incorrect assumptions of the model but also to compensation mechanisms that are not included in it.

Dr. Timmerman solved the problem when she realized that a steady amount of DA-agonist is less difficult to maintain. So she started with the DA-agonist and varied the amount of glutamate in the striatum directly (*cf.* Sec. 8.4, Par. 49). In this case the predicted amplification could be observed.

Lab testing of predicted drug effect

Start: B: $\{a(\text{DA-agonist, striatum}) = \langle +, \text{std}\rangle\} \cup H_{BG} \models$
I: $\{a(\text{glutamate-agonist, striatum}) = \text{inc}\} \rightarrow$ P: $\{f(\text{SNR}) = \text{dec}\}$

Goal: I: $\{a(\text{glutamate-agonist, striatum}) = \text{inc}\} \rightarrow$ O?

Result: I $\rightarrow$ O*: $\{f(\text{SNR}) = \text{dec}\}$

In coming to conclusions about the intervention and observations the data have to be interpreted, described and explained. Conflicts in results are scrutinized when they do not fit expectations (*cf.* Sec. 8.5, Par. 60-66).

Data interpretation

Start: $B \cup H \not\models I \rightarrow O$

Goal: $B \cup H \models I \rightarrow O$

Result: B*, I*, O*

I* Wrong probe location?
Influence of anesthetic?

O* Good signal/noise ratio?
Different cell type with same characteristic?

B* Difference model rat and disease?
Effect by other mechanisms?

The problem is to diagnose the cause of a possible anomaly, A revision of an assumption in I or O is dependent on the type and execution of the particular experiment (*cf.* Sec. 8.4, Par. 50-59). The assumptions in H are the last to go. It is protected by the acknowledgement that it ignores important aspects that might be responsible for anomalous observations.

A lot of simplifications are maintained to conceive experiments and make sense of the data. The problem of finding a relation includes decisions about which properties to manipulate, which to observe, and which to ignore. The phenomenon is made by choosing those properties. So it seems that the assumptions in H are most importantly preserved as a guide for further explorations. This is an important part of the use of theory in experimental research.

When the results of this research where published the problem and its result where reduced in the conclusion to just the goal and the main observed results (*cf.* Sec. 8.5, Par. 67). The consequences for the model of the basal ganglia where reserved for the discussion section. So in summary:

Exploration of drug effect

Start: H_{BG}

Goal: I: {a(DA, striatum} → O?

Results: B: {f(striatum) = std) ∪ H_{BG} ⊭

I: {a(D1, striatum) = inc} → O: {f(SNR) = std}

I: {a(D2, striatum) = inc} → O: {f(SNR) = std}

B: {a(DA-agonist, striatum) = inc} ∪ H_{BG} ⊨

I: {a(glutamate-agonist, striatum) = inc} → O: {f(SNR) = dec}

Summary

In the last subsections I extensively analyzed the structure of example problems in the process of discovery neuropharmacological research. In Table 9.5 I summarize and generalize these examples. This practice shows that testing a new prediction of a hypothesis that explains an earlier observation is only one of many ways of making a discovery. All the different problems I discussed lead to different empirical and conceptual discoveries. In the next section I will go deeper into the process of reasoning in explanation and design.

Problem	Start	Goal	Result	Process
	a. Discovering a new drug effect in the clinic			
Clinical drug treatment	O : {pathologic}	I → W : {normal}	I → O* : {side effects}	Focused intervention Broad observation
Lab exploration of an observed drug effect	I → O : {side effects}	I → O? : {biochemistry}	I → O* : {mechanism}	Focused intervention Focused observation
Explanation of an observed drug effect	I → O	H? ⊨I → O	H* ⊨I → O	Focused explanation
	b. Searching a desired drug effect			
Lab exploration of wished drug function	I : {transmitter}	I → O? : {effect}	I → W* : {effect}	Focused intervention Focused observation
Lab testing for wished drug function	I : {drug}	I → W? : {effect?}	I → O* : {effect}	Focused intervention Focused observation
	c. Designing a new drug			
Rational drug design	I : {drug lead}	H ⊨I? → W : {effect}	I* : {variant}	Focused design
Lab testing for wished drug effect	I : {drug}	I → W? : {effect?}	I → O* : {effect}	Focused intervention Focused observation
Explanation of drug function	I → O	H? ⊨I → O	H* ⊨I → O	Focused explanation
	d. Searching a new drug effect in a drug library			
Lab testing for wished drug effect	I : {given drugs}	I → W? : {effect?}	I → O* : {effect}	Broad intervention Focused observation
	e. Searching a new drug by combinatorial chemistry			
Combinatorial drug design	I : {drug lead}	I? → W : {effect}	I* : {variants}	Broad design

Lab testing for wished drug effect	I : {variants}	$I \rightarrow W?$: {effect?}	$I \rightarrow O^*$: {effect}	Broad intervention Focused observation
Lab exploration of drug structure	$I \rightarrow O$: {effect}	$I \rightarrow O?$: {structure?}	$I \rightarrow O^*$: {structure}	Focused intervention Focused observation
	f. Searching a new drug by computational modeling			
Computational drug design	I, H	$H \models I? \rightarrow W$	$H \models I^* \rightarrow P^*$	Focused prediction Focused design
	g. Testing a new drug			
Clinical drug testing, phase I	I	$I \rightarrow O?$	$I \rightarrow O^*$	Focused intervention Broad observation
Clinical drug testing, phase II	I	$I \rightarrow W?$	$I \rightarrow O^*$	Focused intervention Focused observation
Clinical drug testing, phase III	I	$I \rightarrow W?$	$I \rightarrow O^*$	Focused intervention Broad observation
	h. Exploring a biological system			
Lab exploration of a biological system	I	$I \rightarrow O?$	$I \rightarrow O^*$: {path., normal}	Broad intervention Broad observation
Lab exploration of a drug effect	I	$I \rightarrow O?$:	$I \rightarrow O^*$: {path., normal}	Focused intervention Broad observation
	i. Designing a drug treatment			
Rational drug treatment design	$H \models O$: {pathologic}	$H \models I? \rightarrow W$: {normal}	I^*	Focused design
Clinical drug testing	I	$I \rightarrow W?$	$I \rightarrow O^*$: {side effects}	Focused intervention Broad observation
Design of a treatment of side effect	$H \models I \rightarrow O$: {side effects}	$H \models I? \rightarrow W$	I^*	Focused design
	j. Exploring a drug effect			
Explanation of a drug effect	$I \rightarrow O$	$H? \models I \rightarrow O$	$H^* \models I \rightarrow O$	Focused explanation
Exploration of a drug effect	$H \models I \rightarrow P$	$I \rightarrow O?$	$I \rightarrow O$	Focused prediction Focused observation
Prediction of a drug effect	$H \models I \rightarrow O$	$H \models I \rightarrow P?$	$H \models I \rightarrow P^*$	Focused prediction
Rational design of an experiment condition	$H \models I \rightarrow P$	$H \models I? \rightarrow W$	I^*	Focused design
Lab testing of a predicted drug effect	$H \models I \rightarrow P$: {pred. effect}	$I \rightarrow O?$: {pred. effect?}	$I \rightarrow O^*$	Focused intervention Focused observation
Data interpretation	$B \cup H \not\models I \rightarrow O$	$B \cup H \models I \rightarrow O$	B^*, I^*, O^*	Focused description

Table 9.4: Overview of the structure of discussed problems in drug research.

9.5 Reasoning

In the last section I described the structure of problems in neuropharmacology. We saw that different research activities lead to different kinds of discoveries. Intervention and observation can lead to new empirical discoveries about the natural world. We also saw that reasons to do a particular intervention or observation in a particular way or in a particular location are often suggested by conceptual discoveries that are the result of explanation, prediction and design.

I will now take a closer look at these three kinds of reasoning processes in neuropharmacology. I will discuss logical and computational models of those processes in problem solving that do not specifically aim to explain the cognitive processes that are involved when humans solve these problems, such as is aimed at with ACT-R models. That is a modeling task that requires a different approach. However both descriptive and normative models can share the initial assumptions (start), the goal condition and sometimes the result. In science, these can ideally all be described symbolically. Yet a psychological model will usually differ, compared to a normative computational model, in its modeling of the *process* of solving a problem. The description of the discovery cases in the last section could be a basis for both a psychological ACT-R model, as well as a problem solving model with other constraints.

In this section I will discuss several models of the reasoning processes in scientific problem solving, with a normative pretension. I will argue how these computational models of qualitative explanation, prediction and design could be used to aid the problems in the domain of neuropharmacology, by discussing examples based on the case study.

Explanation and prediction in biology

To formally describe the process of reasoning in explanation and prediction in neuropharmacology I first discuss two computer models of Peter Karp's that model those processes. Karp investigated the development of knowledge about the biological process of attenuation (Karp, 1992; Karp, 1993).

He encoded intermediate states of knowledge about biological objects and processes so that his genetic simulator program GENSIM could use it to simulate experiments and compute predictions. The HYPGENE program takes these predictions as input and compares it with given observations made during an experiment. If there is a discrepancy, HYPGENE modifies assumptions about the initial objects or the processes to explain the difference between GENSIM's prediction and the observation.

Karp considers the process of hypothesis formation as employed by HYPGENE to be a design problem. In this way a hypothesis is an artifact to be synthesized and is subject to design constraints, such as among others consistency with the data, predictive success and simplicity. HYPGENE modifies a theory to satisfy constraints by implementing design rules. Karp derived these from his historical study of knowledge about attenuation, which on its turn provided a test bed for the development of HYPGENE and GENSIM.

Peter Karp's research goal was to model the competence of biologists, not their performance, by identifying reasoning mechanisms that are sufficient to solve hypothesis formation problems in biology, regardless whether they are valid psychologically (Karp, 1992). To implement GENSIM and HYPGENE, he made use of effective tools for search control and an assumption based truth maintenance system.

The program GENSIM can make qualitative predictions about biochemical processes. Types of chemical objects are represented in a taxonomic hierarchy in a frame based Class Knowledge Base (C). Theories about biological processes like chemical reactions are represented as production rules in a Process Knowledge Base (T). The process rules define what classes of objects participate in a reaction and what conditions must be true for the reaction to occur. Process rules further specify what new objects are created if the rule's conditions are met.

GENSIM can predict the outcome of an experiment by applying the process rules to the given specified objects at the start. It is assumed that no objects are entirely consumed during a reaction, and therefore GENSIM only *adds* new objects monotonically to the initial ones. In this way GENSIM can predict what objects should result if the assumptions about initial conditions and the processes are correct. Because GENSIM implements a qualitative chemistry it does not make predictions concerning concentrations or reaction rates. Yet, it can predict increasing and decreasing quantities of chemical compounds.

The HYPGENE program can make qualitative explanations to account for the difference between predictions from GENSIM and observations. As input HYPGENE takes GENSIM's initial conditions I_a, *i.e.* the set of statements about objects that are assumed to be initially present in an experiment that is named "a", the predicted outcome P_a, *i.e.* the set of statements about the objects after the experiment, plus dependency information that records how P_a was computed from I_a, the prediction error $Error_a$, *i.e.* the difference between the prediction and the observation, and also access to all elements in the class knowledge base C and the process knowledge base T. In short, the input contains all elements from the tuple $\langle I_a, P_a, Error_a, T, C\rangle$. P_a is entailed by the initial conditions I_a and the processes in T:

$$I_a \cup T \models P_a$$

An experiment is anomalous when O_a, *i.e.* the set of statements about observation made in an empirical experiment, is not equal to the predicted observation P_a. The prediction error $Error_a$ is by definition $P_a \Delta O_a$, *i.e.* the symmetric difference between prediction P_a and observation O_a (see also Figure 9.3):

$$P_a \Delta O_a := (P_a - O_a) \cup (O_a - P_a)$$

The main goal of HYPGENE is to eliminate $Error_a$. To achieve that goal HYPGENE reasons backwards from the difference between P_a and O_a. Its sub-goals become to remove statements about objects from P_a not in O_a, to modify assumptions about properties of objects in P_a, to modify assumptions about the quantity of objects in P_a, and to add assumptions about objects from O_a that where not in P_a. To achieve these sub-goals two main types of design operator are employed, those that redesign statements in I_a to I_a*, and those that modify statements in T to T* in such a way that:

$$I_a^* \cup T^* \models P_a^* \text{ and } P_a^* = O_a$$

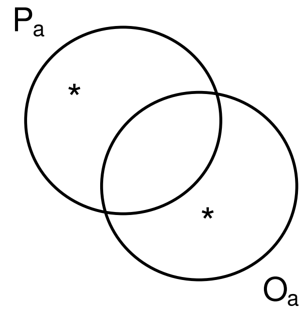

Figure 9.3: The symmetric difference between P_a, the set of statements about objects as predicted, and O_a, the set of statements about the observed objects.

HYPGENE examines the outstanding goals one by one, choosing operators that may satisfy it. For example regard the following simplified GENSIM prediction. For short a statement such as "x" means that there are objects of kind x present in the set-up of experiment *a*. Furthermore we have the following sets: I_a: {x, y}, P_a: {x, y, z} and O_a: {x, y, v}. Say we have process rules r_1, r_2, r_3 as part of T such that:

r_1: = x & y $\rightarrow$ z
r_2: = x & z $\rightarrow$ v
r_3: = x & w $\rightarrow$ v

For example, to remove the assumption about object z from P_a which is not in O_a a design rule can disable process R_1 that causes the metabolism to an object of kind z by modifying its input, assuming that *e.g.* y was actually not an element of I_a and therefore also not in O_a. To explain that a statement about the observation of object v was present in O_a a design rule can modify process rule r_2 assuming that object kind z is not necessary to cause v, or another operator can assume that w was also an element of I_a. So the problem becomes as follows:

GENSIM prediction

Start: I_a: {x, y}, T: {r_1, r_2, r_3}
Goal: $I_a \cup T \models P_a$?
Result: P_a: {x, y, z, v}

HYPGENE explanation

Start: I_a: {x, y} $\cup$ T $\models P_a$: {x, y, z}, O_a: {x, v}
Goal: $I_a? \cup T? \models O_a$
Result: I_a*: {x} $\cup$ T*: {r_1, r_2*: = x $\rightarrow$ v, r_3} $\models O_a$

The representations of objects in C, and the conditions and actions of the processes in T, are much more complex. For the details of the hierarchy of different design rules see (Karp, 1992). There are also operators that modify assumptions about quantity and about the structure of classes C.

It may seem odd to change I_a, but I_a only represents assumptions about what objects are present during an experiment. In biological practice knowledge of initial conditions is often uncertain because of the complexity of objects under study and the sometimes unpredictable effects of laboratory techniques. Karp found that it is normal practice in biology to first take a closer look at the assumed initial conditions, before changing hard earned theories. This practice was confirmed in the case of testing the basal ganglia model. Yet it is possible to slightly revise the model to explain the observed effect of dopamine agonists.

Explanation of a drug effect

The HYPGENE and GENSIM programs model qualitative reasoning in biochemistry So, they are able to model the process of explanation and prediction about transmitter metabolism in the brain. Yet, reasoning in neuropharmacology is also about increasing and decreasing values of variables. Reasoning about these aspects is better modeled by the QSIM program.

As we saw in section 9.3, given a qualitative differential equation, QSIM can make qualitative predictions about the behavior of a dynamical system. Richards et al (1994) devised a program that does the opposite. Given a qualitative description of the behavior of a system the program MISQ infers a qualitative differential equation that implies that behavior. I shall apply the techniques of this program to an example of the Parkinson case.

In the detailed discussion of the exploration of the effect of dopamine we saw that the basal ganglia model predicted an inhibitory effect in the SNR after a dopamine-agonists intervention in the striatum.

QSIM prediction

Start: B: $\{f(\text{striatum}) = \langle 0, \text{std}\rangle\}$, H_{BG}
Goal? $B \cup H_{BG} \models$ I: $\{a(\text{DA-agonist, striatum}) = \text{inc}\} \rightarrow$ P?
Result: P: $\{f(\text{SNR}) = \text{dec}\}$

Yet under basal activation of the striatum the effect was not significant. The prediction error can be traced to constraints C.9 and C.10 of H_{BG}.

C.9 $d/dt\, f(\text{striatum-D1-to-SNR}) = M^{++}(f(\text{striatum}), a(\text{DA, striatum}))$
C.10 $d/dt\, f(\text{striatum-D2-to-GPe}) = M^{+-}(f(\text{striatum}), a(\text{DA, striatum}))$

In C.9 the activity of the direct pathway *f*(striatum-D1-to-SNR) is positively dependent on the activity of the striatum *f*(striatum) and the amount of dopamine *a*(DA, striatum) that can act on the D1-receptor. So an increase of dopamine will cause the same amount of increase of activation of the direct pathway, regardless of the activation of the striatum. The same goes for the inhibition of the indirect pathway as defined in C.9. A program like MISQ is able to suggest different constraints that imply the observed values. The observed effects can be accounted for with a revision of C.9 and C.10 to C.9* and C.10*:

MISQ explanation

Start: B: $\{f(\text{striatum}) = \langle 0, \text{std}\rangle\}$, H_{BG}
I: $\{a(\text{DA-agonist, striatum}) = \text{inc}\} \rightarrow$ O: $\{f(\text{SNR}) = \text{std}\}$
Goal: $H_{BG}? \models I \rightarrow O$
Result: $\{C.9^*, C.10^*\} \in H_{BG}^*$

C.9* $d/dt\, f(\text{striatum-D1-to-SNR}) = f(\text{striatum}) \times a(\text{DA, striatum})$
C.10* $d/dt\, f(\text{striatum-D2-to-GPe}) = f(\text{striatum}) / a(\text{DA, striatum})$

Now if there is only low basal activity of the striatum then DA will have a lot less effect then when the activity of the striatum is higher. Based on the revised hypothesis a new prediction can be deduced. An increase of activation of the striatum together with an increase of dopamine agonists now implies a stronger effect.

Prediction of a drug effect

A formal description of qualitative theories such as the basal ganglia model can also be useful in the research practice itself. The problem of the basal ganglia model, as noted in Chapter 3, is that it is too simple to be real and becomes too complex to work with when it would be extended to incorporate details.

The advantage of a formal description is that you can add more kinds of details, while you can still easily explore predictions by making use of a computer program like QSIM that easily computes the consequences for the variables you are interested in. As an example I list a number of computable predictions of different effects on the SNR after intervening in the direct and indirect pathways of the basal ganglia with selective dopaminergic agonists. Comparing these kinds of predictions with lab observations can result into more detailed and accurate models of biological structures such as of the basal ganglia.

QSIM prediction

Start: B: $\{f(\text{striatum}) = \langle +, \text{std}\rangle\}$, H_{BG}, I
Goal: $H_{BG} \models I \rightarrow P?$

D1- agonists

Start: I: $\{f(\text{striatum-D1-to-SNR}) = \text{inc}, f(\text{striatum-D2-to-GPe}) = \text{inc}\}$
Result: P: $\{a(\text{GABA, SNR}) = ?, a(\text{Glu, SNR}) = \text{inc}, f(\text{SNR})) = \text{inc}?\}$

Start: I: $\{f(\text{striatum-D1-to-SNR}) = \text{std}, f(\text{striatum-D2-to-GPe}) = \text{inc}\}$
Result: P: $\{a(\text{GABA, SNR}) = \text{dec}, a(\text{Glu, SNR}) = \text{inc}, f(\text{SNR})) = \text{inc}\}$

D2-agonists

Start: I: $\{f(\text{striatum-D1-to-SNR}) = \text{dec}, f(\text{striatum-D2-to-GPe}) = \text{dec}\}$
Result: P: $\{a(\text{GABA, SNR}) = ?, a(\text{Glu, SNR}) = \text{dec}, f(\text{SNR})) = ?\text{dec}\}$

Start: I: $\{f(\text{striatum-D1-to-SNR}) = \text{dec}, f(\text{striatum-D2-to-GPe}) = \text{std}\}$
Result: P: $\{a(\text{GABA, SNR}) = \text{dec}, a(\text{Glu, SNR}) = \text{std}, f(\text{SNR})) = \text{inc}\}$

Ratios of D1 and D2 agonists:

Start: I: {*f*(striatum-D1-to-SNR) = std, *f*(striatum-D2-to-GPe) = dec}
Result: P: {*a*(GABA, SNR) = dec, *a*(Glu, SNR) = dec, *f*(SNR)) = ?}

Start: I: {*f*(striatum-D1-to-SNR) = inc, *f*(striatum-D2-to-GPe) = std}
Result: P: {*a*(GABA, SNR) = inc, *a*(Glu, SNR) = std, *f*(SNR)) = dec}

Rational drug design

Another process that is an important part of pharmacology is design. Vos and Kuipers (1992) proposed that the development of design research in general can best be described as a more or less systematic attempt to bring together the properties of available materials and the demands derived from intended applications. They proposed a set-theoretic model of this process. In this model there is a set RP of all relevant properties for a product to be developed. A subset W of RP includes the wished-for properties of the intended product, so RP-W is the set of unwanted properties. For each possible prototype x that is created there is an operational profile, consisting of a set of operational properties O(x) that are part of RP, see Figure 9.4.

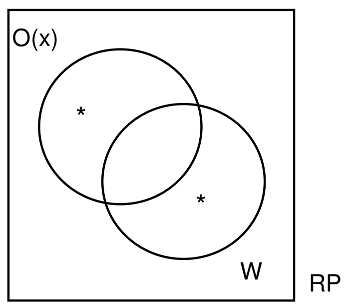

Figure 9.4: The symmetric difference between operational properties O(x) of prototype x, and the set of wished-for properties W, both part of relevant properties RP for the object to be developed.

A problem-state during development can now be described as the symmetric difference W Δ O(x), defined by the set of unrealized wanted properties together with the set of realized properties that are not wanted, *i.e.*:

$$W \,\Delta\, O(x) := (W - O(x)) \cup (O(x) - W)$$

W Δ O(x) denotes the set of problems, *i.e.* the qualitative deviation of O(x) from W. The number of problems to be solved is defined as |W Δ O(x)|, indicating the quantitative deviation.

The goal of design research is to develop a better product x* such that ideally O(x*) is closer to W. Kuipers & Vos propose a descriptive model of the transitions of problem states. This model defines a proper assessment criterion for the improvement of problem state transitions. Prototype x* is an improvement of x in view of W iff:

$$O(x^*) \,\Delta\, W \text{ is a proper subset of } O(x) \,\Delta\, W$$

So a new prototype is an improvement only if it has at least one wished-for property more or at least one unwished-for property less.

For most design research it is possible to divide the set of relevant properties into two complementary sets of structural and functional properties S and F. Often first a functional profile WF is determined of what the product is supposed to do. The next question is how this can be realized. Yet a product is not uniquely determined by WF, often functional equivalents are possible, so the set of looked for structural properties is called an appropriate structural profile AS for WF if it causally implies the desired functional properties WF.

In drug research the determination of the desired functionality WF is normally guided by known characteristics of a disease. These can be explicated as part of the set of conceivable characteristics of potential applications C(A) of a drug, see Figure 9.5. We can say that for a given disease y its profile C(y) uniquely determines the desired functional profile WF, while the reverse is not the case. A drug for a given characteristic can be useful for each disease containing that characteristic. An improvement of a drug's structure and functionality can be defined analogous to the definition above.

For example, in the case of Parkinson's disease, the characteristics of the pathological condition C(Parkinson's disease) includes a degeneration of dopaminergic neurons in the substantia nigra, and a subsequent depletion of the neurotransmitter dopamine in that area. For symptomatic treatment it is suggested that a drug is found that replaces the function of dopamine in that area (WF). This should be an effect caused by a drug with an appropriate structure (AS). How the properties AS of a drug might by suggested by WF is extensively discussed in Vos (1991).

In the next subsection I will discuss how the functional properties WF for a drug intervention can be rationally inferred given characteristics of a disease C(y) and (a formal description of) an available biological theory.

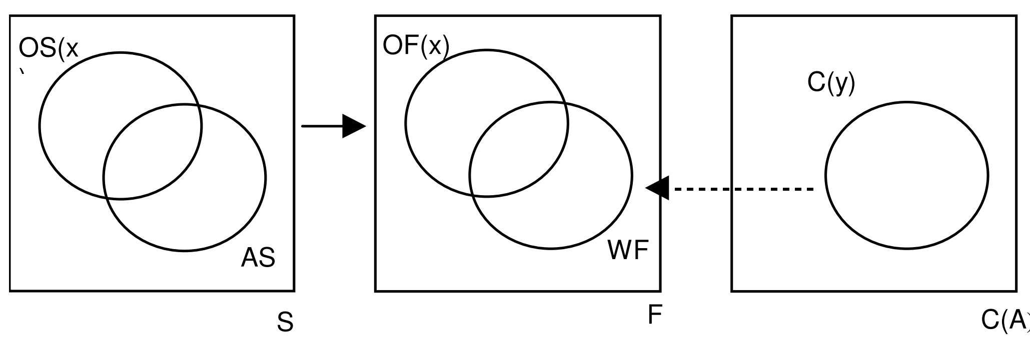

Figure 9.5: A problem state in the S/FA model of drug design research. The disease characteristics C(y) determine the wished-for biochemical effect WF that is caused by a looked for drug with appropriate structure AS. OS(x) and O*F*(x) are the operational structural and functional profile of prototype drug x.

Rational drug treatment design

In section 9.4 we saw that in the rational design of a drug treatment knowledge of biological processes is used to infer the effect of a drug intervention. The suggested intervention can either contain a description of the desired local influence of a drug on the system, or a description of a drug that is known to have the needed functional properties. These desired properties of the drug should cause a decrease in disease symptoms, and are called a drug lead (Vos, 1991). The rational search for a drug lead can be understood as a problem of qualitative reasoning. Knowledge of qualitative relations between variables describing properties of a pathological biological system can be sufficient to find variables that can influence that system.

The search involved is structurally similar to that of explanatory reasoning, but has a different search goal. Instead of finding a simple hypothesis that explains an observed behavior, the task is to find a minimal intervention that has a desired effect on properties such as the behavior of the system, with minimal side effects. So, analogously to inference to the best explanation, this process can be called "inference to the best intervention".

The object of drug treatment design does not initially concern the properties of a compound as in drug design, but the properties of a biological system, an organism. In the latter the goal is to create a drug so that it has given desired properties, in the former the goal is to create the behavior of a biological system so that it has given desired properties. These properties can also be divided in structural and functional properties. A disease is a set of unwanted properties of a biological system. These can be compared with wished for properties of a system. So we can define the characteristics of a disease as follows.

Definition 11 *Disease characteristics.* Given the operational properties O(x) of a pathological system x and the wished for properties W, the characteristics C(y) of a disease y can be defined as the symmetric difference between O(x) and W:

$$C(y) := W \,\Delta\, O(x)$$

The set O(x) contains all the considered properties of a system x, not only the pathological properties. So the set $W \cap O(x)$ is not empty. The goal of drug treatment is to change the properties O(x) of system x to O*(x) such that both $O^*(x) - W$ and $W - O^*(x)$ are minimized

Rational drug treatment design involves finding a drug treatment for a given pathological condition of a system by maximally employing known theories and knowledge about biological processes. A proper theory about a disease should be able to explain the pathological properties.

So, let a set H of theories about biological processes be given as well as background assumptions B(x) involved in the explanation of the observed properties among the properties O(x) of a pathological system x. The problem of the design of a drug treatment of the pathological properties $O(x) \,\Delta\, W$ is to cause only wished for properties from W by a drug intervention I(x) of the system, *i.e.* $H \cup B(x) \models I(x) \rightarrow W$. If we can explain the pathological condition, then we can use that knowledge to infer a suitable intervention.

Rational drug treatment design

Start :	$H \cup B(x) \models O(x)$
Goal :	$H \cup B(x) \models I? \rightarrow W$
Result :	$I^*(x)$

The search goal is to find, by reasoning about processes in H, a proper drug intervention that influences processes that cause the desired properties W, but not those from O(x) – W. That is, the goal is to eliminate the difference between W and O(x). The result of the search is the suggestion of a manipulation of a local biochemical property that can be affected by a drug. A drug that has this wished for functional effect (WF) can be searched for in the set of known drugs, or pose a new problem for rational drug design.

Of course it would be ideal, given the known H and the nature of the disease, to infer a suggestion for a drug intervention I that only causes W. A drug usually also causes side effects, often creating undesired effects that are not part of the disease that is targeted. Therefore we need a gradual evaluation criterion for the improvement of suggestions (*cf.* T.A.F. Kuipers, Vos en Sie 1992). Let us say that the moderated design goal is to find the suggestion I such that its (predicted) consequence for a system $H \cup B(x) \models I(x) \rightarrow P(x)$ resembles the desired condition W more than the pathological condition O(x), *i.e.* that:

$P(x) \Delta W$ is a proper subset of $O(x) \Delta W$

That is, roughly, the drug should not have more unwanted consequences than accomplished desired consequences, *cf.* Figure 9.6.

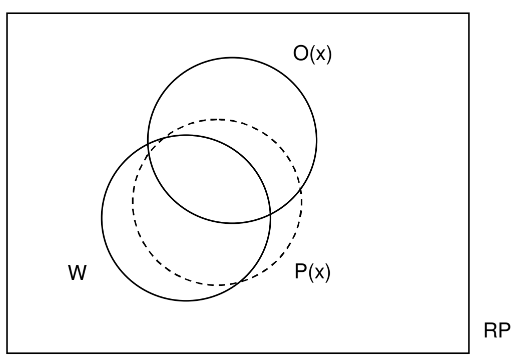

Figure 9.6: Problem state in searching an intervention with effect P(x) that most resembles desired properties W in treating a pathological system x with operational properties O(x).

The evaluation of improvement of more than one drug suggestion can follow the same lines. A drug intervention I* of x is better than an intervention I if the properties of consequence P* resemble W more than those of P:

$P^*(x) \Delta W$ is a proper subset of $P(x) \Delta W$

However, this is only an evaluation of properties that is neutral to the different kinds of undesired properties. In this way an intervention could be inferred that treats most of the symptoms, but causes a symptom that is worse than the disease that is treated. This could be remedied by a ordering of the undesired properties, together with a quantitative measure of deviation.

The resulting suggestion for a drug intervention can on its turn be used to test the theories used to find the suggestion. Given an inferred drug intervention I(x), an experiment can be done and its resulting observation of the altered operational properties O(x) of x can be compared with the predicted properties P(x). A discrepancy can be used to redesign H, or the assumptions about B(x) or I(x).

Testing of predicted drug effect

Start: $H \cup B(x) \models I(x) \rightarrow P(x)$
Goal: $I \rightarrow O?$
Result: $I \rightarrow O^*$

The same kind of design and testing is found in designing experimental conditions for focused testing of hypotheses, as we saw at the end of section 9.4.

Rational design of an experimental condition

Start : $H \cup B \models I \rightarrow P$
Goal : $H \cup B \models I? \rightarrow W$
Result : I^*

Computational drug lead discovery

Next to rational drug design in the lab, we also saw in section 9.4 that there is a subdiscipline in pharmacology called computational drug design. This discipline is concerned with the rational design and exploration of drug structures and drug function, making use of computational models of those structures. This is usually a quantitative approach, making massive computations in quantum mechanics to predict *e.g.* the folding of the protein structure of a receptor in reaction to a drug structure.

To search for a drug treatment there are many kinds of computer programs that can help to diagnose a particular disease and suggest a drug treatment. These programs make use of explicitly known established assumptions about pathology and medicine. These kinds of tools are less known in the practice of basic research.

An exception is the ARROWSMITH program of Swanson and Smallheimer (1997). This program searches for unknown relations between research findings in the literature. One research group the members of which know each other's writings may establish that there is a connection between biological properties A and B, while another group in a slightly different field could have established a link between B and C, but may be unaware of the other group's results. If C is related to a pathological property then A might be a lead candidate for drug treatment. The ARROWSMITH program searches for these implicit links in the published texts that describe results, making use of different statistical techniques.

A search by ARROWSMITH discovered a link between fish oil (A) and Reynaud's disease (C). Both are related to properties of blood viscosity, platelet aggregation, and vascular reactivity (B). The program also discovered a relation between magnesium and migraine, they share 11 related properties. Weeber *et al* (2000b) develop techniques for the same kind of problems in order to find out whether published side effects of drugs may be beneficial for the treatment of other diseases.

Textual drug lead discovery

Start: …, A → B, …, B → C, …
Goal: ? → C
Result: A*

This search for a novel treatment is conducted in the enormous amount of published results, represented by the dots in the above scheme A discovered relation is a discovery of an implicit conceptual relation in explicitly known results. These results contain explicit descriptions of interventions, observations, explanations and predictions. Such searches are fruitful, but they can not find implicit consequences of proposed explanations and theories. This is a hard problem because a description of a theory in natural or informal language, as is common in medicine, is difficult to analyze semantically. It is possible to cluster words that are related in meaning, but with current techniques it is not possible to computationally infer logical consequences from sentences in natural language. These techniques can assist in the discovery of textual relations, but they still have to be interpreted by a knowledgeable scientist.

A more formal description of both qualitative and quantitative results can result in computational discoveries of new interesting consequences, for both basic research and treatment.

Logical drug lead discovery

Start: $H \models C$
Goal: $H \models ? \rightarrow C$
Result: $H \models A^* \rightarrow C$

For example the formal description of the basal ganglia, incorporating more research details can be used to computationally explore interesting predictions, as we saw in this section, and to search interventions with desired properties in detail. The desired state W that should be caused by an intervention aimed to combat Parkinson's disease includes {*f*(SNR) = dec}. A search program can infer that this can be caused by an increase of GABA or a decrease of glutamate: {*a*(GABA, SNR) = inc, *a*(Glu, SNR) = dec}. These variables can on their turn be influenced by different interventions with selective dopamine agonist. The effects of those where predicted earlier in this section. We already saw in section 9.3 how the model implied the traditional treatment of targeting the metabolism of dopamine.

Computational drug lead discovery

Start: H_{BG}
Goal: QSIM(H, I?) = W
Result: I*

In this search not only one looks for which variables are related, but also one tries to find out how the values of these variables influence each other. It is one thing to know that dopamine and Parkinson symptoms are related, but it is a more specific hypothesis that the decrease of the amount of dopamine is related to the increase of symptoms. With this knowledge available in QDEs one may better evaluate possible interventions in a system. If an intervention causes a variable to be steady while you wished that it would increase, then that result is not as bad as when it would decrease, *cf.* Table 9.5. This evaluation can be extended with specific weights for particular properties, depending on the disease and the importance of particular properties.

P \W	dec	std	inc
dec	1	0	-1
std	0	1	0
inc	-1	0	1

Table 9.5: Quantitative evaluation of a predicted qualitative value of a variable from the set of predicted properties P, compared with the desired value of the same variable in the set of wished-for properties W.

In this way all variables under consideration can be evaluated to find the best intervention, for which the sum of the evaluations of all value comparisons should be maximal. This evaluation can also be applied to the other structurally similar rational design problems I discussed.

9.6 Conclusion

So, what is the rational use of theory and experiment in the process of scientific discovery in the practice of drug research for Parkinson's disease, compared to theory as discussed in part II? In Part II we saw that in the theoretical conception of scientific discovery the rational use of a theory in scientific discovery is to explain observations with simple additional hypotheses that can predict properties of phenomena. The rational use of experiment is to test predictions of those explanations. This use of theory and experiment is considered to be what makes the process of scientific discovery rational.

After analyzing a practice of scientific discovery in detail it is apparent that both theory and experiments are used for many more reasons, with different goals and results, all leading to different kinds of scientific discoveries, as summarized in Table 9.4. Theories are used to explain observations, and to make predictions that test the theory. But they are also used to rationally design experimental conditions and treatments, and to explore phenomena. Experiments are used to test theories by observing and intervening in properties of phenomena that are predicted by that theory. But they are also used in treatment and in many different kinds of exploration, often intervening in ways and having one look in to directions that are not suggested by theory, but just by curiosity. In this case the rational to use experiment to explore those areas is that there *is* no theory or expectation about it, giving ample room for new empirical discoveries.

Yet devising and testing theory remains an important part of science, both in theory and in practice. The use and nature of theories in scientific practice is grounded in primary and secondary cognitive mechanisms as explicated in Chapter 5. Natural language and informal diagrams are the vehicles of choice to represent assumptions in the scientific discipline I analyzed. Yet in this way it is not always possible to fully oversee the consequences of known theories and results. To better understand these, epistemologists devise formal theories about theories. These can be used to represent a theory more explicitly. In the case study, I showed how this can be done for the theoretical model of the basal ganglia.

So, it is now time to answer the specific questions of this thesis for my case study:

Question 1 The theory of the basal ganglia consists of qualitative relations between variables of chemical and electrical neural activity in nuclei. The structure of this theory can be represented as a qualitative differential equation. In a structuralist approach the theory can be defined by its models, given a set of constraints on conceptually possible models defined by a set of variables and possible values.

Question 2 When a drug intervention is observed together with a certain change in a property of a biological system, a conditional dependency can be inferred. Clusters of observed or assumed relations between variables that describe the domain can together become a theory that explains other relations. A goal in neuropharmacology is to infer a hypothesis H that best explains how observed properties O of a biological system are conditionally dependent on an intervention I, *i.e.* to infer the best explanation (IBE). This hypothesis can be used to rationally design a drug treatment or a condition for an experiment, to cause wanted properties W, *i.e.* to infer the best intervention (IBI). Given a hypothesis and an intervention new consequences P can be predicted. If the goal is to test the theory, the goal of the reasoning process is to infer the best prediction (IBP), see Table 9.6.

Problem	Start	Background	Process	Goal	Goal properties
Explanation	$I \rightarrow O$	B, V, Q, D	IBE	H*	$B \cup H^* \models I \rightarrow O$
Design	W	B, H: $\langle V, Q, C, D\rangle$	IBI	I*	$B \cup H \models I^* \rightarrow W$
Prediction	I	B, H: $\langle V, Q, C, D\rangle$	IBP	P*	$B \cup H \models I \rightarrow P^*$

Table 9.6: Main processes of reasoning discussed in this part.

Finding a hypothesis H*, an intervention I* or a prediction P* that is optimal given the assumed conditions, can be called a conceptual discovery and is often no trivial problem. It may require an exhaustive search in a problem space that is defined out of known concepts in V and Q. Within that problem space a hypothesis H may be found that is denoted by the set of constraints C on all the possible models of a domain, as determined by V, Q and D. An intervention and prediction are rationally searched for within the models allowed by the constraints of the hypothesis. Finding a proper hypothesis for a domain may also require a conceptual revision of the problem space, by revising the variables in V and quantity spaces in Q. In contrast to a conceptual discovery, making an experimental intervention and observation can lead to an empirical discovery, when new properties or phenomena are actually created or observed.

Question 3 In this chapter I have discussed many different routes between theory and experiment. Particular interventions and observations can lead to new empirical discoveries when the observed properties are not expected, or prove an expectation to be wrong. Such an empirical discovery is able to logically refute a theory, but in practice it will not be deserted. A false theory can remain a fruitful pointer to directions for new interventions and observations that can lead to new empirical discoveries. In biological practice explanations are revised to fit observations, looking first at the assumptions about the interventions and observations and in the background.

In contrast to the discussed diversity in the discovery process in the practice of neuropharmacology, I end my discussion of the questions of this thesis with a formal summary of the textbook example of the process of discovery in drug research for Parkinson's disease:

1. Observe phenomenon p: pi,…,pj (parts of the basal ganglia)
2. Describe p: $I \rightarrow O$
   I: $\{a(\text{DA, striatum}) = \text{dec}\}$
   $I \rightarrow O$: $\{f(\text{SNR}) = \text{inc}\}$
3. Explain p: $B \cup H? \models I \rightarrow P$
   $H_{BG}*$: $\langle V, Q, C, D\rangle$
4. Predict p: $B \cup H_{BG} \models I \rightarrow P?$
   I: $\{a(\text{DA, striatum}) = \text{inc}\} \rightarrow P*$: $\{f(\text{SNR}) = \text{dec}\}$
   Design p: $B \cup H_{BG} \models I? \rightarrow W$
   I*: $\{a(D_{1/2}\text{-agonist, striatum}) = \text{inc}\} \rightarrow W$: $\{f(\text{SNR}) = \text{dec}\}$
5. Intervene p: do I
6. Observe p: see P?

In step 1. processes and properties of the basal ganglia are observed. It is described how a decrease of dopamine in the striatum by an intervention results into an increase of activity of the SNR. A model of the basal ganglia is proposed that implies the observation in step 3. In step 4. this model is used to predict that an increase of dopamine in the striatum will cause a decrease of activation of the SNR. Given the decrease of SNR activation as a wished-for property, the model also implies other possible interventions, such as agonists for a receptor-subtype. These suggestion can be experimentally tested in steps 5. and 6.

This process can be aided in both theory and practice with the use of computer modeling tools that can assist in finding descriptions, explanations, predictions, and new designs. However, the bigger problem to make these tools useful is the availability of biological theory in a formal representation. It would be ideal if scientists in biology would publish their results both in natural language and in a formal format. To this end, Peter Karp started an internet database that invites biologists to add their results in a provided formal format. This database is used to test new methods that can aid the process of discovery in science, aiding on its turn the process of understanding rationality in discovery.

*

Summary

Part I Introduction

The specific problem adressed in this thesis is: what is the rational use of theory and experiment in the process of scientific discovery, in theory and in the practice of drug research for Parkinson's disease? The thesis aims to answer the following specific questions: what is: 1) the structure of a theory?; 2) the process of scientific reasoning?; 3) the route between theory and experiment? In the first part I further discuss issues about rationality in science as introduction to part II, and I present an overview of my case-study of neuropharmacology, for which I interviewed researchers from the Groningen Pharmacy Department, as an introduction to part III.

Part II Discovery

In this part I discuss three theoretical models of scientific discovery according to studies in the fields of Logic, Cognition, and Computation. In those fields the structure of a theory is respectively explicated as: a set of sentences; a set of associated memory chunks; and as a computer program that can generate the observed data. Rationality in discovery is characterized by: finding axioms that imply observation sentences; heuristic search for a hypothesis, as part of problem solving, by applying memory chunks and production rules that represent skill; and finding the shortest program that generates the data, respectively. I further argue that reasoning in discovery includes logical fallacies, which are neccesary to introduce new hypotheses. I also argue that, while human subjects often make errors in hypothesis evaluation tasks from a logical perspective, these evaluations are rational given a probabilistic interpretation.

Part III Neuropharmacology

In this last part I discusses my case-study and a model of discovery in a practice of drug research for Parkinson's disease. I discuss the dopamine theory of Parkinson's disease and model its structure as a qualitative differential equation. Then I discuss the use and reasons for particular experiments to both test a drug and explore the function of the brain. I describe different kinds of problems in drug research leading to a discovery. Based on that description I distinguish three kinds of reasoning tasks in discovery, inference to: the best explanation, the best prediction and the best intervention. I further demonstrate how a part of reasoning in neuropharmacology can be computationally modeled as qualitative reasoning, and aided by a computer supported discovery system

Propositions

1. Problem
Before one can stand on the shoulders of giants, one first has to climb them.

2. Rationality
Assumptions about processes of scientific discovery imply assumptions about psychological processes, and *vice versa*.

3. Neuropharmacology
A part of reasoning in neuropharmacology can be modeled as reasoning about qualitative differential equations, and can be assisted by a computer.

4. Logic
Reasoning in scientific discovery includes logical fallacies, which are necessary to introduce new hypotheses by abduction.

5. Cognition
To understand the rationality of (secondary) cognitive processes of symbolic problem solving in science, one also needs to understand how these processes are controlled by (primary) cognitive processes of probabilistic learning.

6. Computation
One has learned something when one can compute part of the same output with less input.

7. Theory
Rationality in discovery, in theory, includes inferring hypotheses that best explain observations, and inferring predictions that can experimentally test those hypotheses best.

8. Practice
Rationality in discovery, in practice, also includes inferring the best interventions in designing drugs, treatments, and experimental conditions to explore phenomena.

9. Discovery
Interdisciplinary scientists build bridges that other scientists are not eager to cross.

Samenvatting

Deel I Introductie

Het probleem dat dit proefschrift behandeld is: wat is het rationeel gebruik van theorie en experiment in het proces van wetenschappelijk ontdekken, zowel in theorie als in de praktijk van geneesmiddelenonderzoek voor de ziekte van Parkinson? Een antwoord wordt gegeven op de volgende specifieke vragen, wat is: 1) de structuur van een theorie; 2) het proces van wetenschappelijk redeneren; en 3) de route tussen theorie en experiment? In deel I behandel ik verder, als introductie voor deel II en III, debatten over rationaliteit in wetenschap, en presenteer ik een overzicht van mijn casestudie van de neurofarmacologie, waarvoor ik enkele onderzoekers van het Universitair Centrum voor Farmacie van de Rijksuniversiteit Groningen interviewde.

Deel II Ontdekken

In dit deel behandel ik drie modellen van ontdekken, volgens de logica, de cognitieve psychologie, en de computerwetenschap. In deze velden wordt een theorie respectievelijk gezien als: een verzameling zinnen; een verzameling geassocieerde geheugenpartjes; en, een computerprogramma. Rationaliteit in ontdekken is gekarakteriseerd als respectievelijk: het vinden van axioma's waaruit observatiezinnen afgeleid kunnen worden; het heuristisch zoeken naar een verklaring voor observaties waarbij geheugenpartjes en regels, die vaardigheden representeren, worden toegepast; en, het vinden van het kortste computerprogramma die de observatiedata kan genereren. Ik beargumenteer dat redeneringen die nieuwe hypothesen introduceren de vorm hebben van een drogreden. Verder beargumenteer ik dat, terwijl proefpersonen vaak incorrecte evaluaties van hypothesen maken vanuit een logisch perspectief, deze evaluaties rationeel zijn vanuit een probabilistische interpretatie.

Deel III Neurofarmacologie

In dit laatste deel behandel ik mijn casestudie van ontdekken in een praktijk van geneesmiddelenonderzoek voor de ziekte van Parkinson. Ik bespreek de dopamine theorie van de ziekte van Parkinson en modelleer de structuur. Daarna behandel ik het gebruik van experimenten om zowel een medicijn te testen als om functies van de hersenen te verkennen. Ik beschrijf daarbij verschillende soorten problemen in geneesmiddelenonderzoek die leiden tot een ontdekking. Gebaseerd op deze beschrijving maak ik een onderscheid tussen drie soorten redeneertaken in wetenschappelijk ontdekken, het afleiden van: de beste verklaring, de beste voorspelling, en de beste interventie. Ik demonstreer verder hoe een deel van het redeneren in de neurofarmacologie kan worden gemodelleerd als kwalitatief redeneren, en kan worden ondersteund door de computer.

Stellingen

1. Probleem
Voordat men kan staan op de schouders van reuzen, moet men deze eerst beklimmen.

2. Rationaliteit
Aannames over processen van wetenschappelijk ontdekken impliceren aannames over psychologische processen, en *vice versa*.

3. Neurofarmacologie
Een deel van het redeneren in de neurofarmacologie kan worden gemodelleerd als redeneren over kwalitatieve differentiaalvergelijkingen, en kan worden geassisteerd door een computer.

4. Logica
Drogredenen zijn een deel van het redeneren in wetenschappelijk ontdekken, dit is noodzakelijk voor het introduceren van nieuwe hypothesen door abductie.

5. Cognitie
Om de rationaliteit van (secundaire) cognitieve processen in het oplossen van symbolische wetenschappelijke problemen te begrijpen, is het ook nodig om te begrijpen hoe deze processen worden gestuurd door (primaire) cognitieve processen van probabilistisch leren.

6. Computatie
Je hebt iets geleerd als je een deel van dezelfde output kunt berekenen met minder input.

7. Theorie
Rationaliteit in ontdekken, in theorie, omvat het afleiden van hypothesen die het best observaties verklaren, en het afleiden van experimenten die deze hypothesen het best testen.

8. Praktijk
Rationaliteit in ontdekken, in de praktijk, omvat tevens het afleiden van de beste interventies in het ontwerpen van medicijnen, behandelingen, en experimentele condities om fenomenen te verkennen.

9. Ontdekken
Interdisciplinaire wetenschappers bouwen bruggen die ander wetenschappers niet graag oversteken.

Bibliography

Aliseda-LLera, Atocha (1997) *Seeking Explanations: Abduction in Logic, Philosophy of Science and Artificial Intelligence*. ILLC dissertation series DS-1997-04, Amsterdam.

Anderson, J. R. (1985) *Cognitive psychology and its implications*. Freeman, New York,

Anderson, J. R. (1993) *Rules of the Mind*. Lawrence Erlbaum Associates.

Anderson, J. R. (1996) ACT, A Simple Theory of Complex Cognition. American Psychologist, 51(4), 355-365.

Anderson, J. R. & C. Lebiere (1998) *The Atomic Components of Thought.* Lawrence Erlbaum Associates.

Bechtel, W. (1988) *Philosophy of Science, an overview for cognitive science*. Lawrence Erlbaum Associates.

Cooper, J. R., Bloom, F.E. & Roth, R.H. (1996*) The Biochemical Basis of Neuropharmacology* (*7th ed.)* Oxford: Oxford University Press.

Culp, S. & Kitcher, P. (1989) Theory Structure and Theory Change in Contemporary Molecular Biology. Britisch Journal of Philosophy of Science(40), 459-483.

Cuto, L. & Crutcher, M.D. (1991) The Basal Ganglia. In E.R. Kandel, J.H. Schwartz & T.M. Jessel (Eds.), Principles of Neural Science (pp. 647-659, Ch. 42) New Jersey: Prentice Hall.

D.R. Swanson, N.R. Smalheiser (1997) An interactive system for finding complementary literatures: a stimulus to scientific discovery. AI, 91, pp.183-203.

Darden, L. (1990) Diagnosing and fixing faults in theories. In: J. Shrager & P. Langley, o.c., 319-354.

Darden, L. (1991) *Theory change in science: strategies from Mendelian genetics*. Oxford University Press, New York.

Darden, L. (1997) Recent Work in Computational Scientific Discovery. In M.G. Shafto & P. Langley (Ed.), 19th Annual Conference of the Cognitive Science Society, (pp. 161-166) Stanford: Lawrence Erlbaum Associates.

de Jong, H. & Rip, A. (1997) The computer revolution in science: steps towards the realization of computer supported discovery environments. Artificial Intelligence(91), 225-256.

de Jong, H. & van Raalte, F. (1997) Comparative Analysis of Structurally Different Dynamical Systems. In M.E. Pollack (Eds.), Proceedings of the 15th International Joint Conference on Artificial Intelligence (pp. 486-491) San Francisco: Morgan Kaufmann.

de Jong, H. (1998) *Computer-supported analysis of scientific measurements*. PhD-thesis, Twente University.

de Jong, H., A. Rip (1997a) The computer revolution in science: steps toward the realization of computer-supported discovery environments. Artificial Intelligence, 91 pp. 225--256.

de Jong, H., Mars, N.J.I. & van der Vet, P. (1996) CEC: Comparative analysis by envisionment construction. In W. Wahlster (Eds.), Proceedings of the 12th European Conference on Artrificial Intelligence (pp. 476-480) Chichester: John Wiley and Sons.

de Jong, H., Mars, N.J.I. & van der Vet, P.E. (1998) Computer-supported analysis of measurements in materials science. In J. Meheus (Ed.), International Congress on Discovery and Creativity, Gent: University of Gent.

Carnap, R. (1967) *The logical structure of the world.* (R.A. George trans.) Berkeley, University of California Press. (originally published in 1928.)

De Vries, G. (1985) *De ontwikkeling van wetenschap, Een inleiding in de wetenschapsfilosofie.* Wolters-Noordhoff.

DeJong, M.R. (1990) Primate models of movement disorders of basal ganglia origin. TINS, 13(7), 281-285.

Dunbar, K. (1995) How Scientists Really Reason: Scientific Reasoning in Real-World Laboratories. In R.J. Sternberg & J.E. Davidson (Eds.), The Nature of Insight, Oxford, MIT Press, pp. 365--395.

Flach, P. (1995) *Conjectures, an inquiry concerning the logic of induction*. ITK Dissertation series 1995-1, University of Tilburg.

Fodor, J.A. (1975) *The Language of Thought*. Thomas Y. Cromwell Company.

Forbus, K. (1984) Qualitative process theory. Artificial Intelligence, 24: 85-168.

Forbus, K.D. (1997) Qualitative Reasoning. In A.B. Tucker (Eds.), The computer science and engineering handbook (pp. 715-733) Boca Rato, Florida: CRC Press.

Giere, R. (1981) *Understanding scientific reasoning* 2^{nd} *edition*. Holt, Rineman and Winston, New York.

Giere, R. (ed.) (1992) *Cognitive models of science*. University of Minnesota, Minneapolis.

Giere, R. N. (1987) Cognitive models in the philosophy of science. In A. Fine & P. Machander (eds.), PSA-1986, Philosophy of Science Association, East Lansing.

Giere, R. N. (1988) *Explaining science: A cognitive approach*. The University of Chigago Press, Chigago.

Gingerich, O. (1992) *The great Copernicus chase and other adventures in astronomical history* Cambridge, Mass.: Sky ; Cambridge : Cambridge University Press.

Goldman, A. (1986) *Epistemology and Cognition.* Harvard University Press.

Gorman, M. (1992) *Simulating science*. Indiana university press, Bloomington.

Hacking, I. (1983) *Representing and intervening*. Cambridge University Press.

Hendriks, P., Taatgen, N. & Andringa, T. (1997*) Breinmakers en breinbrekers, Inleiding Cognitiewetenschap*. Amsterdam: Addison Wesley Longman.

Holland, J.H., K.J. Holyoak, R.E. Nisbett, P.R. Thagard (1986) *Induction: processes of inference, learning, and discovery*. MIT-Press, Cambridge.

Horn, A.S. (1990) Dopamine Receptors. In J.C. Emmet (Eds.), Membranes & receptors (pp. 229-290) Oxford: Pergamon Press.

Houk, J.C., Davis, J.L. & Beiser, G.B. (1995) Models of Information Processing in the Basal Ganglia. Cambridge, MA: The MIT Press.

Hunter, L. (Ed.) (1993) *Artificial Intelligence and Molecular Biology*. Menlo park, California: MIT Press.

Iwasaki, Y. & Simon, H.A. (1994) Causality and model abstraction. Artificial intelligence : an international journal, 67(1), 143-194.

Jenner, P. (1995) The Rationale for the use of dopamine agonists in Parkinson's disease. Neurology, 45(Suppl. 3), S6-S12.

Kamps, J. (2000) *A Logical Approach to Computational Theory Building* (*with applications to sociology*. Phd-thesis, ILLC dissertation series 2000-02, Amsterdam.

Kandel, E.R. (1991) Disorders of Thought: Schizophrenia. E.R. Kandel, J.H. Schwartz & T.M. Jessel (Eds.), Principles of Neural Science (pp. 853-868, Ch. 55) New Jersey: Prentice Hall.

Kandel, E.R. (1991) Nerve Cells and Behaviour. E.R. Kandel, J.H. Schwartz & T.M. Jessel (Eds.), Principles of Neural Science (pp. 18-32, Ch. 2) New Jersey: Prentice Hall. Koller, W.C., Silver, D.E. & Lieberman, A.

Kandel, E.R., Schwartz, J.H. & Jessel, T.M. (Ed.) (1991) *Principles of Neural Science* (*3rd ed.)* New Jersey: Prentice Hall.

Kandel, E.R., Siegelbaum, S.A. & Schwartz, J.H. (1991) Synaptic Transmission. In E.R. Kandel, J.H. Schwartz & T.M. Jessel (Eds.), Principles of Neural Science (pp. 123-134, Ch. 9) New Jersey: Prentice Hall.

Karp P. and M. Riley (1993) Representations of metabolic knowledge. In Proceedings of the First International Conference on Intelligent Systems for Molecular Biology, L. Hunter, D. Searls, and J. Shavlik (eds.), AAAI Press , Menlo Park, CA, pp. 207--215.

Karp, P.D. & Mavrovouniotis, M.L. (1994) Representing, Analyzing, and Synthesizing Biochemical Pathways. IEEE Expert: intelligent systems and their applications, 9(2), pp. 11-22.

Karp, P.D. (1992) Hypothesis Formation as Design. In J. Shrager & P. Langley (Eds.), Computational Methods of Scientific Discovery and Theory Formation (pp. 275-317) Palo Alto: Morgan Kaufmann Publishers, Inc.

Karp, P.D. (1993) Design Methods for Scientific Hypothesis Formation and Their Application to Molecular Biology. Machine Learning, 12, pp. 89-116.

Keppel Hesselink¸ J.M. (1986) *De ziekte van Parkinson*. Kerckebosch, Zeist.

Kitcher, P. & Culp, S. (1989) Theory Structure and Theory Change in Contemporary Molecular Biology. British Journal of Philosophy of Science, 40, 459-483.

Koetter, R. & Wickens, J. (1995) Interactions of Glutamate and Dopamine in a Computational Model of the Striatum. Journal of Computational Neuroscience(2), 195-214.

Koetter, R. & Wickens, J. (1998) Striatal mechanisms in Parkinson's disease: new insights from computer modeling. Artificial intelligence in medicine, 13(1), 37-56.

Kuhn, T.S. (1970) *The Structure of Scientific Revolutions*. University of Chicago Press.

Kuipers, B. (1994) *Qualitative Reasoning, Modeling and simulation with incomplete knowledge*. Cambridge, MA: MIT Press.

Kuipers, B. & Kassirer, J.P. (1984) Causal reasoning in medicine: analysis of a protocol. Cognitive Science(8), 363-385.

Kuipers, Benjamin, A. Moskowitz, J.P. Kassirer (1990) Critical Decisions under Uncertainty: Representation and structure.

Kuipers, T.A.F en A.R. Mackor (1995) *Cognitive Patterns in Science and Common Sense*. Rodopi.

Kuipers, T.A.F. (2000) *Structures in Science, heuristic patterns based on cognitive structures*. Kluwer.

Kuipers, T.A.F., Vos, R. & Sie, H. (1992) Design Research Programs and the Logic of their Development. Erkenntnis(37), 37-63.

Kuipers, Th.A.F (1999) Abduction aiming at empirical progress or even truth approximation leading to a challenge for computational modelling. In Foundation of Science 4, 307-323.

Lakatos, I. (1978) *The methodology of scientific research programs*. Cambridge.

Langley, P., H.A. Simon, C.L. Bradshaw, J.M. Zytkow (1987) *Scientific Discovery, Computational Explorations of the Creative Processes.* MIT Press, Cambridge, Massachusetts.

Laudan, L. (1978) *Progress and its problems*. University of California Press.

Li, M. and P.M.B. Vitányi (1990) Kolmogorov Complexity and its Applications. In: J. van Leeuwen Algorithms and Complexity, Handbook of Theoretical Computer Science, volume A, Elsevier,187--254.

Li, M. and P.M.B. Vitányi (1993) Inductive Reasoning and Kolmogorov Complexity. In: Journal of Computer and System Sciences, vol. 44, no. 2.

Li, M. and P.M.B. Vitányi (1994*) An Introduction to Kolmogorov Complexity and Its Applications*. Addison-Wesley, Reading, MA.

Marr, D. (1982) *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W.H. Freeman and Company, NY.

Mitchell, I.J., Brotchie, J.M., Brown, G.D.A. & Crossman, A.R. (1991) *Modeling the functional Organization of the Basal Ganglia, A Parallel Distributed Processing Approach.* Movement Disorders, 6(3), 189-204.

Newell, A. and H. Simon (1972) *Human Problem Solving*. Englewood Cliffs, N.J.: Prentice Hall.

Oaksford, M, N. Chater (1996) Rational Explanation of the Selection Task In: Psychological review, ISSN 0033-295X, Vol. 103 (Issue 2), pp. 381-391 (11)

Parent, A. & Cicchetti, F. (1998) The Current Model of Basal Ganglia Organization Under Scrutiny. Movement Disorders, 13(2), 199-202.

Popper, K. R. (1959) *The logic of scientific discovery*. Hutchinson.

Richards, B.L., Kraan, I. & Kuipers, b.J. (1991) Automatic Abduction of Qualitative Models. In Proceedings of the 5th International Workshop on Qualitative Reasoning about Physical Systems (pp. 295-301)

Rikken, F. (1998) *Adverse drug reactions in a different context: A scientometric approach towards adverse drug reactions as a trigger for the development of new drugs*. PhD-thesis, Groningen University.

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 & Philosophy, 11: 63-80.

Schaffner, K.F. (1987) Computerized implementation of biomedical theory structure: An artificial intelligence approach. A. Fine & P. Machander (eds.), PSA-1986, Philosophy of Science Association, East Lansing.

Schaffner, K.F. (1993) *Discovery & explanation in biology & medicine*. University of Chicago Press, Chicago.

Schaffner, K.F. (ed.) (1985) *Logic of discovery and diagnosis in medicine*. University of California Press, Berkeley.

Schwartz, J.H. (1991) Chemical messengers: small molecules and peptides. In E.R. Kandel, J.H. Schwartz & T.M. Jessel (Eds.), Principles of Neural Science (pp. 213-224, Ch. 14) New Jersey: Prentice Hall.

Shrager J. & P. Langley (eds.) (1990) *Computational models of scientific discovery and theory formation*. Kaufmann, San Mateo.

Simon, H.A. (1996) *The Sciences of the Artificial* (*3rd ed.)* Cambridge: MIT Press.

Solomonoff, R.J. (1964) A Formal Theory of Inductive Inference, part I. In Information and Control, 7, 1--22.

Swanson, D.R. & Smalheiser, N.R. (1997) An interactive system for finding complementary literatures: a stimulus to scientific discovery. Artificial Intelligence(91), 183-203.

Taatgen, N.A.. (1999). *Learning without limits: from problem solving toward a unified theory of learning*. PhD Thesis, Groningen University.

Thagard, P. & Verbeurgt, K. (1998) Coherence as Constraint Satisfaction. Cognitive science : a multidisciplinary journal of artificial intelligence, psychology, and language, 22(1), 1-24.

Thagard, P. (1988) *Computational Philosophy of Science*. MIT-Press, Cambridge.

Thagard, P. (1992) *Conceptual revolutions*. Princeton University Press.

Thagard, P. (1996) The concept of disease: structure and change. Communication and cognition : a quarterly journal, 29(3-4), 445-478.

Thagard, P. (1998) Explaining Disease: Correlations, Causes, and Mechanisms. Minds and machines : journal for artificial intelligence, philosophy, and cognitive science, 8(1), 61-78.

Thagard, P. (1999) *How Scientists Explain Disease*. Princeton Universty Press, New Jersey.

Timmerman, W. (1992) Dopaminergic receptor agents and the basal ganglia : pharmacological properties and interactions with the GABA-ergic system. PhD-thesis, Groningen University.

Timmerman, W. (1992) *Dopaminergic receptor agents and the basal ganglia : pharmacological properties and interactions with the GABA-ergic system.* PhD-thesis. Groningen University.

Timmerman, W., F. Westerhof, T. van der Wal, B.C. Westerink: 1998, Striatal dopamine-glutamate interactions reflected in substantia nigra reticulata firing. Neuroreport 9, pp. 3829--3836.

Valdés-Pérez, R.E. (1998) Why Some Machines do Science Well. In working notes of the International Congress on Discovery and Creativity, Ghent.

van den Bosch, A.P.M. (Ed.) (1994) *De Bestorming van het Brein. Leren en adapteren in hersenen, geest en computer*. Groningen: Studium Generale Groningen.

van den Bosch, A.P.M. (1994) *Computing Simplicity, About the role of simplicity in discovery, explanation, and prediction*. Master thesis, Department of Philosophy, University of Groningen.

van den Bosch, A.P.M. (1994) Philosophical analysis of the research into anti-Parkinson medicines with the aid of computational models. In T.A.F. Kuipers & M. Ter Hark (Eds.), Aard en achtergrond multi- en interdisciplinair onderzoek in gedrags-, cognitie- en neurowetenschappen (pp. 53-59) Groningen: Department of Philosophy, section WLK.

van den Bosch, A.P.M. (1995) Discovering Patterns by Searching for Simplicity. In R. Valdez-Perez (Eds.), Systematic Methods of Scientific Discovery. Papers from the 1995 AAAI Spring Symposium (pp. 166-171) Menlo Park, California: The AAAI Press.

van den Bosch, A.P.M. (1996) Abductieve Inferentie als primaire cognitie. In G. Groot, H. Oosterling & A. Prins (Ed.), Van agora tot markt : acta van de 18e Nederlands-Vlaamse Filosofiedag, Rotterdam: Faculteit der Wijsbegeerte van de Erasmus Universiteit Rotterdam.

van den Bosch, A.P.M. (1996a) Learning Abductive Search by Analogy in ACT-R. In J. Van den Herik & T. Weijters (Ed.), BENELEARN-96, (pp. 179-188) Maastricht: MATRIKS/ Universiteit Maastricht.

van den Bosch, A.P.M. (1996b) Modeling Scientific Discovery in ACT-R. In J.A. Anderson (Ed.), Third annual ACT-R Workshop Proceedings. Pittsburgh: Department of Psychology, CMU.

van den Bosch, A.P.M. (1997) Rational Drug Design as Hypothesis Formation. In P. Weingartner, G. Schurz & G. Dorn (Ed.), 20th International Wittgenstein Symposium, I (pp. 102-108) Kirchberg am Wechsel (Aus): The Austrian Ludwig Wittgenstein Society.

van den Bosch, A.P.M. (1998) Qualitative Drug Lead Discovery. In working notes of the International Congress on Discovery and Creativity, Ghent, pp. 163--165.

van den Bosch, A.P.M. (1999) Inference to the Best Manipulation - a case study of qualitative reasoning in neuropharmacy. In Foundations of Science 4 (4). Special issue on Scientific Discovery and Creativity: Case studies and computational approaches. Guest editors: J. Meheus & T. Nickles. p. 483-495.

Van Eemeren, F. H., & R. Grootendorst. (1992) *Argumentation, communication and fallacies*. Lawrence Erlbaum, Hillsdale N.J.

Verhagen-Kamerbeek, W.D.J. (1994) *Noradrenergic and Dopaminergic Therapy in Parkinson's Disease.* PhD-Thesis, University of Groningen.

Vermeulen, R.J. (1994) *Effects of Dopamine D1 and D2 receptor agonists on motor behavior of MPTP-lesioned monkeys*. PhD-thesis, Vrije Universiteit Amsterdam.

Vos, R. (1991) *Drugs looking for diseases._Innovative drug research and the development of the beta blockers and the calcium antagonists*. Kluwer Academic Press, Dordrecht.

Waterson, B. (1988) *Weirdos from another planet.* Andrews & McMeel, Kansas.

Weeber, M. (2000b) *Literature-based Discovery in Biomedicine*, PhD-thesis, Groningen University.

Weeber, Marc, Henny Klein, Lolkje T. W. de Jong-van den Berg, and Rein Vos (2000a) Text-based discovery in biomedicine: The architecture of the *DA*D-system. In *Proceedings of the 2000 AMIA. Annual Fall Symposiu*m. Hanley and Belfus, Philadelphia, PA.

Wichmann, T. & DeLong, M.R. (1993) Pathophysiology of Parkinsonian Motor Abnormalities. Advances in Neurology, 60, 53-61.

Index

Titles in the ILLC Dissertation Series

ILLC DS-1996-01: **Lex Hendriks**
Computations in Propositional Logic

ILLC DS-1996-02: **Angelo Montanari**
Metric and Layered Temporal Logic for Time Granularity

ILLC DS-1996-03: **Martin H. van den Berg**
Some Aspects of the Internal Structure of Discourse: the Dynamics of Nominal Anaphora

ILLC DS-1996-04: **Jeroen Bruggeman**
Formalizing Organizational Ecology

ILLC DS-1997-01: **Ronald Cramer**
Modular Design of Secure yet Practical Cryptographic Protocols

ILLC DS-1997-02: **Natasa Rakic**
Common Sense Time and Special Relativity

ILLC DS-1997-03: **Arthur Nieuwendijk**
On Logic. Inquiries into the Justification of Deduction

ILLC DS-1997-04: **Atocha Aliseda-Llera**
Seeking Explanations: Abduction in Logic, Philosophy of Science and Artificial Intelligence

ILLC DS-1997-05: **Harry Stein**
The Fiber and the Fabric: An Inquiry into Wittgenstein's Views on Rule-Following and Linguistic Normativity

ILLC DS-1997-06: **Leonie Bosveld - de Smet**
On Mass and Plural Quantification. The Case of French 'des'/'du'-NP's

ILLC DS-1998-01: **Sebastiaan A. Terwijn**
Computability and Measure

ILLC DS-1998-02: **Sjoerd D. Zwart**
Approach to the Truth: Verisimilitude and Truthlikeness

ILLC DS-1998-03: **Peter Grünwald**
The Minimum Description Length Principle and Reasoning under Uncertainty

ILLC DS-1998-04: **Giovanna d'Agostino**
Modal Logic and Non-Well-Founded Set Theory: Translation, Bisimulation, Interpolation

ILLC DS-1998-05: **Mehdi Dastani**
Languages of Perception

ILLC DS-1999-01: **Jelle Gerbrandy**
Bisimulations on Planet Kripke

ILLC DS-1999-02: **Khalil Sima'an**
Learning efficient disambiguation

ILLC DS-1999-03: **Jaap Maat**
Philosophical Languages in the Seventeenth Century: Dalgarno, Wilkins, Leibniz

ILLC DS-1999-04: **Barbara Terhal**
Quantum Algorithms and Quantum Entanglement

ILLC DS-2000-01: **Renata Wasserman**
Resource Bounded Belief Revision

ILLC DS-2000-02: **Jaap Kamps**
A Logical Approach to Computational Theory Building (with applications to sociology)

ILLC DS-2000-03: **Marco Vervoort**
Games, Walks and Grammars: Problems I've Worked On

ILLC DS-2000-04: **Paul van Ulsen**
E.W. Beth als logicus

ILLC DS-2000-05: **Carlos Areces**
Logic Engineering. The Case of Description and Hybrid Logics

ILLC DS-2000-06: **Hans van Ditmarsch**
Knowledge Games

ILLC DS-2000-07: **Egbert L.J. Fortuin**
Polysemy or monosemy: Interpretation of the imperative and the dative-infinitive construction in Russian

ILLC DS-2001-01: **Maria Aloni**
Quantification under Conceptual Covers

ILLC DS-2001-02: **Alexander P.M. van den Bosch**
Rationality in Discovery: A Study of Logic, Cognition, Computation and Neuropharmacology